AD________________

Award Number: DAMD17-01-1-0328

TITLE: Computer-Aided Characterization of Breast Masses on Volumetric Ultrasound Images: An Adjunct to Mammography

PRINCIPAL INVESTIGATOR: Berkman Sahiner, M.D., Ph.D.

CONTRACTING ORGANIZATION: University of Michigan Medical Center
Ann Arbor, Michigan 48109-0331

REPORT DATE: October 2004

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
October 2004

3. REPORT TYPE AND DATES COVERED
Annual (6 Sep 2003 - 5 Sep 2004)

4. TITLE AND SUBTITLE
Computer-Aided Characterization of Breast Masses on Volumetric Ultrasound Images: An Adjunct to Mammography

5. FUNDING NUMBERS
DAMD17-01-1-0328

6. AUTHOR(S)
Berkman Sahiner, M.D., Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Michigan Medical Center
Ann Arbor, Michigan 48109-0331

E-Mail: berki@umich.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

The purpose of this project is to develop computer vision techniques for the analysis of sonographic images of breast masses, and to combine computerized sonographic and mammographic analyses. The techniques developed in this project are aimed at providing a second opinion to the radiologists in the task of making a biopsy recommendation. In the second year of the project, we have (1) compared the accuracy of the classifier designed in the first year of this project to that of experienced radiologists; (2) conducted studies on the effect of the developed classifier on radiologists' characterization of breast masses on ultrasound images; and (3) investigated methods for combining computer classification methods based on ultrasound and mammogram images. Our results indicate that the accuracy of our computer classifier is similar to that of experienced breast radiologists on our data set. We have also shown that experienced radiologists can significantly ($p<0.006$) improve their mass characterization accuracy on sonograms when aided by our algorithm. Our results on combining computer classification methods based on ultrasound and mammogram images indicate that multi-modality computer-aided diagnosis may further improve the classification accuracy.

14. SUBJECT TERMS
Computer-aided diagnosis; ultrasonography; breast masses; breast cancer detection

15. NUMBER OF PAGES
93

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	***18. SECURITY CLASSIFICATION OF THIS PAGE***	***19. SECURITY CLASSIFICATION OF ABSTRACT***	***20. LIMITATION OF ABSTRACT***
Unclassified	Unclassified	Unclassified	Unlimited

(3) Table of Contents

(4) Introduction

At present, biopsy is the gold standard in breast lesion characterization. However, the positive breast biopsy rate is only about 15-30%. This means that 70-85% of breast biopsies are performed for benign lesions. In order to reduce patient anxiety and morbidity, as well as to decrease health care costs, it is desirable to reduce the number of benign biopsies without missing malignancies. Mammography and sonography are two low-cost imaging modalities that may be improved so that radiologists can obtain more accurate diagnostic information to differentiate malignant and benign lesions. Computerized analysis of the lesions on these images is one of the promising tools that may improve the radiologists' accuracy in characterizing these lesions by providing a consistent and reliable second opinion to radiologists.

In this project, our goal is to analyze volumetric images to improve the accuracy of computerized sonographic breast lesion characterization, and to combine these characterization results with those obtained by computerized analysis of mammograms. Computerized image analysis, feature extraction, and classification methods will be developed to characterize breast masses on three-dimensional or volumetric ultrasound (US) images. The output of the classifier will be a computer rating related to the likelihood of malignancy of the mass. The accuracy of this rating will be studied by comparing it to the biopsy results. We will then combine this rating with a similar rating obtained by computerized analysis of the mammograms of the same patient. The combined classifier is expected to be more accurate than either classifier alone.

(5) Body

In the current project year (9/6/03-9/5/04), we have performed the following studies:

(A) Collection of Database

We have continued the collection of database in the third year of this proposal. Up to this point in time, we have collected volumetric data from 183 patients. We have digitized over 215 mammograms from over 85 patients where each case contained a biopsy proven mass. The breast imaging experts in this project, Drs. Helvie and Roubidoux, have continued reading mammograms and US volumes, marking the masses and providing mass descriptors.

We have also performed an experiment in which six expert radiologists provided mass descriptors of 102 US scans. The assessment of mass characteristics shown in Table 1 helped us better identify the properties of our data set.

Overall US impression	Shape	Margins	Echogenicity	Through transmission	Other features
Negative B: 0 (0) M: 0 (0)	Oval B: 32 (70) M: 13 (23)	Circumscribed B: 27 (59) M: 3 (5)	Echogenic B: 0 (0) M: 0 (0)	Increased transmission B: 13 (28) M: 15 (27)	Taller than wide B: 1 (2) M: 11 (20)
Simple Cyst B: 1 (2) M: 0 (0)	Round B: 8 (17) M: 3 (5)	Spiculated B: 1 (2) M: 7 (13)	Isoechoic B: 5 (11) M: 3 (5)	Distal shadowing B: 11 (24) M: 20 (36)	Thin echogenic rim B: 2 (4) M: 1 (2)
Complex Cyst B: 4 (9) M: 1 (2)	Lobulated B: 2 (4) M: 7 (13)	Microlobulated B: 5 (11) M: 20 (36)	Hypoechoic B: 35 (76) M: 36 (64)	Neither B: 22 (48) M: 21 (38)	Ductal extension B: 0 (0) M: 3 (5)
Solid B: 41 (89) M: 55 (98)	Irregular B: 4 (9) M: 33 (59)	Ill defined B: 13 (28) M: 26 (46)	Markedly hypoechoic B: 4 (9) M: 9 (16)		Calcifications B: 1 (2) M: 14 (25)
			Anechoic B: 1 (2) M: 1 (2)		Echogenic halo B: 1 (2) M: 2 (4)
			Heterogeneous B: 1 (2) M: 7 (13)		

Table 1: Characteristics of the sonographic masses in our data set of 102 cases. Each characteristic was determined from the assessments by the six radiologists using a majority voting method, in which the descriptor that was selected by the largest number of radiologists was chosen. The numbers in parentheses are the percentages of the descriptors relative to the total number of benign and malignant masses in the data set. Benign (B): N=46, Malignant (M): N=56

(B) Outcome analysis of the observer performance study that uses the computer classifier designed on 3D US images

In year 2 of the project, we had conducted an observer study to investigate if our computer classifier that uses 3D US volumes would improve radiologists' accuracy in differentiation of malignant and benign breast masses on ultrasound images. In year 3, we conducted an in-depth analysis of the results of this observer performance study. The results of this analysis have been submitted to the journal Radiology as an original research paper [1]. (Appendix 3).

The data set this study consisted of 102 US scans containing 102 biopsy-proven breast masses, of which 46 were benign and 56 were malignant. Our previously developed computer algorithm, explained in detail in Appendix 1 [2], had an area under the ROC curve of A_z=0.92. Five radiologists (RAD1-RAD5) participated as observers. They read the 3D US images using a specially-developed software first without CAD and then with CAD. They provided a likelihood of malignancy (LM) rating under both conditions. For details about the study design, please refer to Appendix 3.

The LM ratings of the radiologists with and without CAD were analyzed using ROC methodology. The area under the ROC curve, A_z, and the partial area index above a sensitivity of 0.9, $A_z^{(0.9)}$ [3] were used as the accuracy measures. For the group of five radiologists, the significance of the change in accuracy with CAD was tested using the Dorfman-Berbaum-Metz

(DBM) multi-reader multi-case (MRMC) methodology [4]. The sensitivity and specificity of each radiologist with and without CAD were compared using an LM rating of 2% as the threshold above which biopsy would be recommended [5, 6].

Table 2 shows the area A_z under ROC curve, and the partial area index $A_z^{(0.9)}$ above a sensitivity of 0.9, for the characterization of the masses in the data set without and with CAD by the 5 radiologists. It is observed that every radiologist showed improvement in both measures when they read with CAD. The improvement was statistically significant for every radiologist. The average ROC curves of the five radiologists with and without CAD, and the ROC curve of the computer classifier are shown in Fig. 1. The improvement was statistically significant as measured by DMB analysis (p=0.006).

Rad.	A_z			$A_z^{(0.9)}$		
No	No CAD	With CAD	p value	No CAD	With CAD	p value
1	0.83±0.04	0.89±0.03	0.0008	0.25±0.10	0.35±0.14	0.17
2	0.81±0.04	0.86±0.04	0.0005	0.14±0.08	0.23±0.12	0.13
3	0.87±0.03	0.91±0.03	0.0486	0.39±0.12	0.53±0.12	0.0747
4	0.82±0.04	0.93±0.02	0.0004	0.39±0.10	0.68±0.09	0.0008
5	0.83±0.04	0.90±0.03	0.0007	0.29±0.10	0.42±0.12	0.0323

Table 2: The area A_z under ROC curve, and the partial area index $A_z^{(0.9)}$ above a sensitivity of 0.9, for the characterization of the masses in the data set without and with CAD by the 5 radiologists. The p value for each radiologist is also shown.

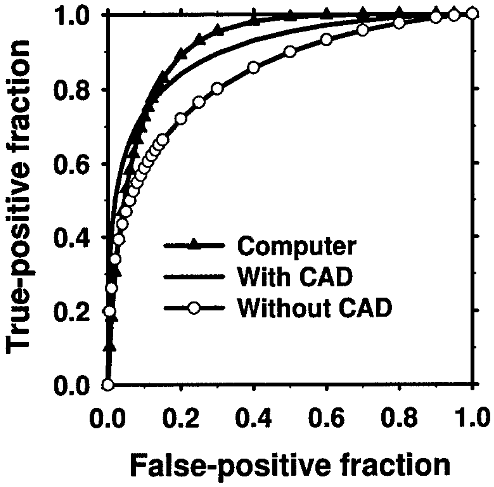


Figure 1: The average ROC curves of the radiologists with and without CAD, and the ROC curve of the computer classifier. With CAD, the average A_z value of the radiologists improved significantly (p=0.006) from 0.84 to 0.90.

With 102 cases and five radiologists, we had a total of 510 pairs of LM ratings with and without CAD. Figure 2 shows a histogram of the change in the radiologists' LM ratings with CAD for these 510 readings. The radiologists did not change their LM rating substantially (i.e., within 5) with CAD in 64% (326/510) of the readings. For malignant masses, the ratings were substantially increased for 34% (95/280) and decreased for 7% (19/280) of the readings. For benign masses, the ratings were substantially increased for 14% (32/230) and decreased for 17% (38/230) of the readings. For benign masses, the decrease in the average LM rating was 0.77, which did not achieve statistical significance (two-tailed p=0.51). The increase in the average LM rating of malignant masses was 5.59, which was statistically significant (two-tailed p<0.0001). Since the "cost" of failing to biopsy a malignant lesion is much greater than that of a benign biopsy, it can logically be expected that radiologists may tend to use the CAD system to confirm and increase their

LM estimate of malignant lesions while not easily reducing the LM estimate of low suspicion lesions. This will result in an overall increase in radiologists' LM ratings. Figure 2 suggests that this is indeed the case in our study.

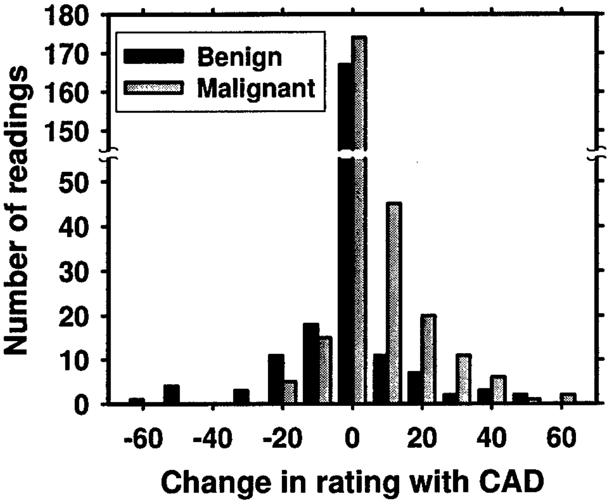


Figure 2: The histogram of the change in radiologists' ratings with CAD. For the majority of the masses (59% of malignant masses and 70% of benign masses) the change was in the range of -4 to 4. When the change in the scores with CAD was greater than or equal to the range of -5 to 5, the change was called substantial. For malignant masses, the ratings were substantially increased for an average of 34% (95/280) and decreased for 7% (19/280) of the readings. For benign masses, the ratings were substantially increased for 14% (32/230) and decreased for 17% (38/230) of the readings.

The sensitivity and specificity of each radiologist with and without CAD at an LM threshold of 2% are listed in Table 3. On the average, the radiologists' sensitivity increased from 96% to 98% with CAD, at the cost of a decrease in specificity from 22% to 19%. The effect of CAD was therefore mixed when measured in terms of the radiologists' sensitivity and specificity values at the threshold of biopsy recommendation (LM of 2%). Table 4 also shows the sensitivity and specificity for each radiologist if the LM threshold were to be adjusted to 7% when they read with CAD, for which the average sensitivity would remain at 96% (same as that without CAD) while the average specificity would increase to 46%. This implies that by appropriate training, it may be possible to translate the benefits with CAD into biopsy decisions that surpass unaided reading in terms of both sensitivity and specificity, or an improvement in specificity without reducing sensitivity.

Rad.	Sensitivity			Specificity		
No.	No CAD*	With CAD*	With CAD**	No CAD*	With CAD*	With CAD**
1	56 (100)	56 (100)	56 (100)	4 (9)	5 (11)	15 (33)
2	51 (91)	53 (95)	49 (88)	12 (26)	11 (24)	28 (61)
3	52 (93)	54 (96)	53 (95)	24 (52)	22 (48)	29 (63)
4	55 (98)	56 (100)	56 (100)	9 (20)	5 (11)	23 (50)
5	56 (100)	56 (100)	56 (100)	1 (2)	1 (2)	11 (24)
Avg.	54 (96)	55 (98)	54 (96)	10 (22)	9 (19)	21 (46)

Table 3.

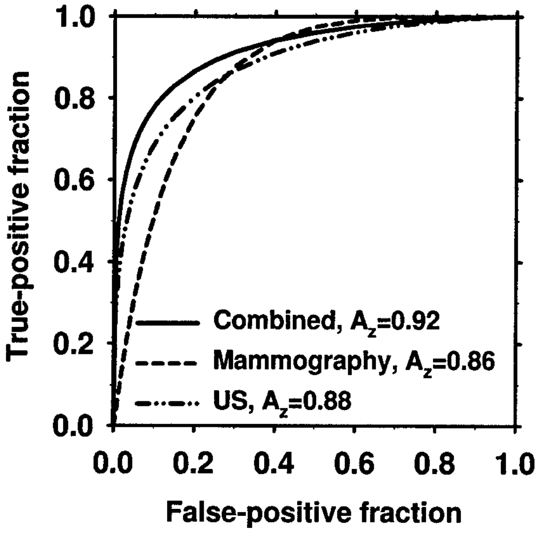
The sensitivity and specificity for each radiologist. In each entry, the first number denotes the number of correctly classified lesions, and the number in parentheses denotes the percentage (i.e., sensitivity for the first three columns. and the specificity for the last three columns). The total numbers of malignant and benign lesions are 56 and 46, respectively. The columns entitled "No CAD*" and "With CAD*" show the sensitivity and specificity at the decision threshold of 2% likelihood of malignancy, without and with CAD, respectively. The columns entitled "With CAD**" show the hypothetical sensitivity and specificity with CAD at a decision threshold of 7% likelihood of malignancy, for which the average sensitivity would be the same as that without CAD (96%), but the average specificity would be increased to 46%.

(C) Further development of the multi-modality classifier

In year 2 of the proposal, we had performed a preliminary study for combining the mammographic and sonographic computerized mass characterization methods. In year 3, we have enlarged our data set for this analysis and performed a more complete study. The results of these studies have been presented at two conferences in 2004 [7, 8](Appendices 2 and 5). We have identified that Method B mentioned in last year's report was the most robust method for combining the classifiers from the two modalities. In this method, we first combine the feature vectors from different mammographic views of the same patient into a case-based mammographic feature vector. Similarly, the feature vectors from different US slices are combined into a case-based US feature vector. The case-based US and mammographic feature vectors are combined into a malignancy score using a single classifier.

We performed a study using US volumes and mammograms from 67 patients to validate this method. Thirty two of the masses were benign and 35 were malignant. The total number of mammographic views was 163, with each case containing between one and three views (CC, MLO, or LAT). Five radiologists read the mammograms and 3D US images on a high-quality computer monitor using a graphical user interface with which they could view the mammographic regions of interest, navigate through 3D volumes, adjust the window and level of the displayed images, and enter a malignancy rating between 1 and 100 (higher rating indicating higher likelihood of malignancy).

The computer classifier using the US images alone, mammograms alone, and the combined feature space had A_z values of 0.88±0.04, 0.86±0.05, and 0.92±0.03, respectively. The ROC curves with the single-modality classifiers and the multi-modality classifier are shown in Figure 3. Although the multi-modality classifier had higher accuracy, the difference between the A_z values of the multi-modality and single-modality classifiers did not reach statistical significance, probably because of the small sample size. The A_z values of the five radiologists ranged between 0.86 and 0.96. The A_z value of their average ROC curve, computed by averaging the a and b values in ROC analysis, was 0.92. Figure 4 compares the ROC curve of the computer classifier to that of the individual radiologists.



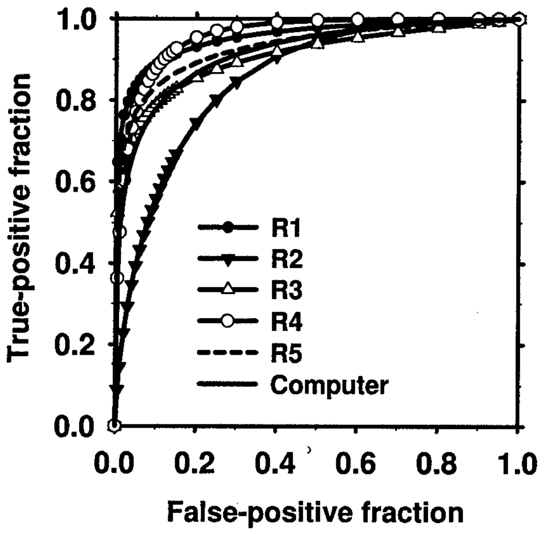


Figure 3: Comparison of the ROC curves for the single-modality and multi-modality computer classifiers.

Figure 4: Comparison of the ROC curves of the five radiologists and the multi-modality computer classifier.

(D) Preliminary work on the effect of the multi-modality classifier on radiologists' classification accuracy

We have been conducting a preliminary study on the effect of the multi-modality classifier described in Section (C) on radiologists' characterization accuracy of breast masses. We have submitted an abstract to RSNA 2004 Conference, and we will present our findings in

December 2004. Up to this point, five experienced radiologists have completed the observer performance study. The data set was the same as that used in Section (C). Using a specially-designed graphical user interface (GUI), the radiologists read the cases sequentially under three conditions: (1) Mammograms alone, (2) Mammograms and 3D US volumes, (3) Mammograms, 3D US volumes, and CAD scores. Under each condition, they provided a likelihood of malignancy rating. They also provided a BI-RADS rating under condition (1) and an action category (follow-up or biopsy) under conditions (2) and (3). Figure 3 shows a snapshot from the GUI under condition (3).

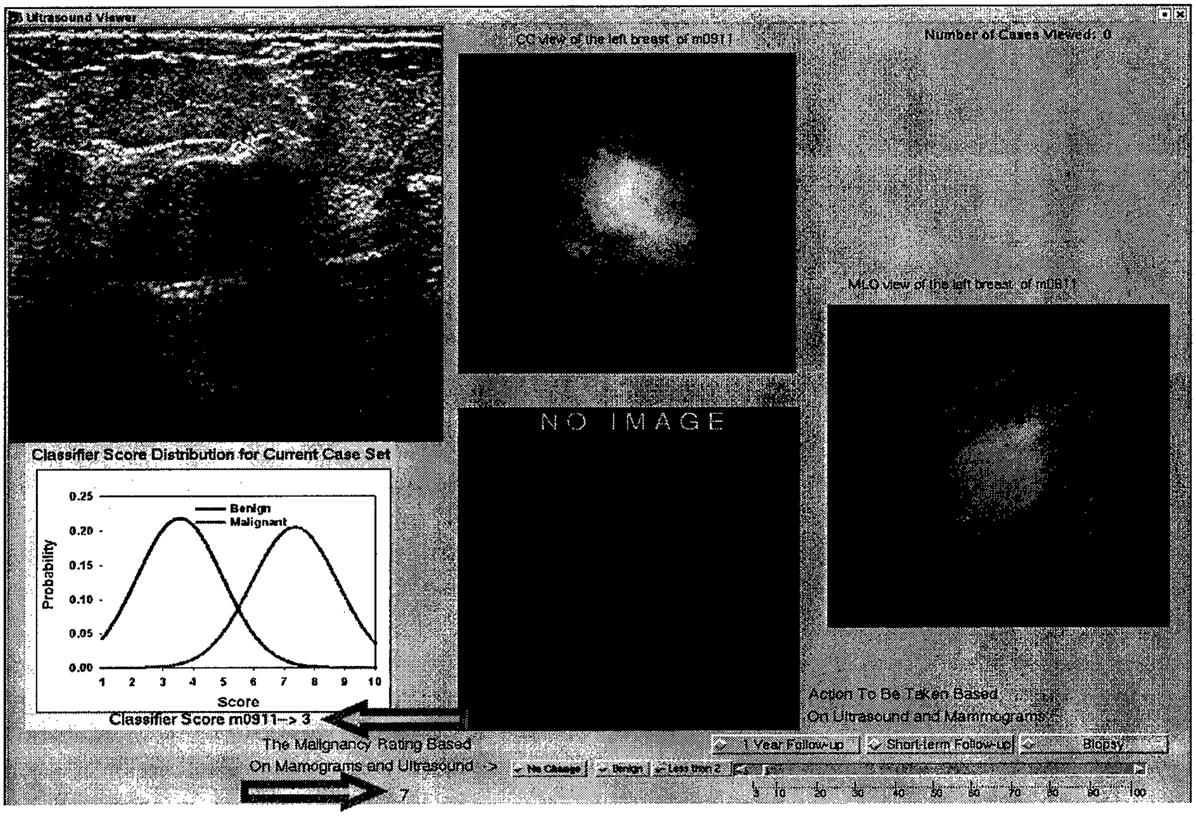

Ultrasound Viewer
CC view of the left breast of m0911
Number of Cases Viewed: 0
MLO view of the left breast of m0911
NO IMAGE
Classifier Score Distribution for Current Case Set
0.25
0.20
0.15
0.10
0.05
0.00
Probability
Benign
Malignant
1
2
3
4
5
6
7
8
9
10
Score
Classifier Score m0911--> 3
The Malignancy Rating Based
On Mamograms and Ultrasound ->
7
No Change
Benign
Less than 2
3
10
20
30
40
50
60
70
80
90
100
Action To Be Taken Based
On Ultrasound and Mammograms
1 Year Follow-up
Short-term Follow-up
Biopsy

Figure 5: A snapshot of the GUI when the radiologist is reading the case under condition 3. The radiologists likelihood of malignancy rating (100 point scale) and the computer score (10-point scale) are shown by arrows.

The classification accuracies of the radiologists under the three reading conditions were quantified using the area under ROC curve, A_z, as well as specificity and sensitivity. Figure 4 shows the A_z values of each radiologist under the three reading conditions. It is observed that for each radiologist condition 3 (reading with CAD) was the most accurate, followed by condition (2) and then followed by condition (1). The average A_z values of the five radiologists were 0.88, 0.92, and 0.95, under Conditions (1), (2), and (3), respectively.

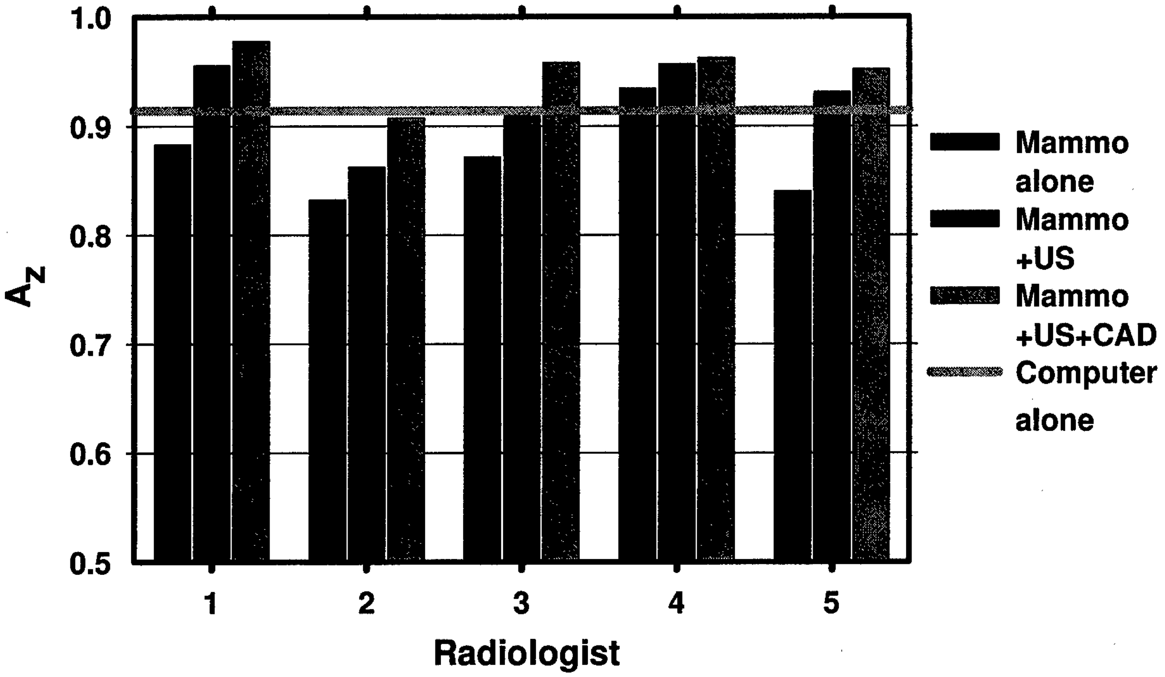


Figure 6: The area A_z under the ROC curve for each radiologist under the three reading conditions. The average A_z values were: 0.88, 0.92, and 0.95, respectively, under Conditions (1), (2), and (3).

The improvement with CAD was statistically significant using the t-test (p=0.03). However, it did not reach statistical significance using the DBM method, likely due to the small sample size. We are continuing to recruit radiologists into the observer study to reduce the variance component in the analysis due to reader variation.

The sensitivities and specificities of the five radiologists are shown in Table 4. It is observed that the radiologists' sensitivity increased when US was used as an adjunct to mammography. However, the trade-off is that the specificity was reduced under condition (2) compared to condition (1). This is not surprising, because these masses were mostly solid. As a result, the radiologists' likelihood of malignancy increased for most masses when they evaluated the case with US. Under condition (3), both sensitivity and specificity increased compared to condition (2), although the increases were small. We expect to obtain a more realistic assessment when more radiologists complete the study.

	Mammography alone		Mammography +US		Mammography +US+CAD	
Rad. #	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
1	0.97	0.34	1.00	0.16	1.00	0.16
2	0.83	0.66	0.89	0.66	0.94	0.63
3	1.00	0.13	1.00	0.13	1.00	0.19
4	1.00	0.16	1.00	0.16	1.00	0.28
5	0.77	0.84	0.94	0.53	0.97	0.50
Avg.	0.91	0.43	0.97	0.33	0.98	0.35

Table 4: The sensitivity and specificity of each radiologist under the three reading conditions.

(6) Key Research Accomplishments

- We performed a through analysis of the observer performance study conducted in year 2 (radiologists reading 3D US volumes without and with CAD). Our analysis confirmed the preliminary finding that CAD can significantly improve radiologists' characterization accuracy of sonographic breast masses.
- Our analysis of the same observer study also implies that by appropriate training, it may be possible to translate the benefits of reading US images with CAD into biopsy decisions that

surpass unaided reading in terms of both sensitivity and specificity, or an improvement in specificity without reducing sensitivity.

- We evaluated the classification accuracy of mammographic CAD (Task 4a), and compared it to the accuracy of US CAD (Task 3).
- We identified preferred algorithms and parameters for the design of the computer classifier that combines the ultrasound and mammography information (Task 3b)
- We performed preliminary studies for the evaluation of the potential improvement when the radiologists use multi-modality CAD (part of Task 4b).

(7) Reportable Outcomes

The journal paper that we submitted to Medical Physics in year 2 has been published in year 3. We have submitted a manuscript to the journal Radiology on the effect of the 3D US classifier on radiologists' characterization of breast masses on ultrasound images. In year 3, we have submitted one conference abstract on computer-aided characterization of breast masses on ultrasound images that has been accepted for publication. Additionally, we presented our results in three conferences (RSNA 2003, SPIE 2004, and IWDM 2004), and published two conference proceeding papers. We are in the process of writing a manuscript for journal submission on the effect of the multi-modality classifier on radiologists' characterization of breast masses on US images and mammograms.

Journal Publications:

Sahiner B, Chan HP, Roubidoux MA, Helvie MA, Hadjiiski LM, Ramachandran A, LeCarpentier GL, Nees A, Paramagul C, Blane CE, "Computerized characterization of breast masses on 3-D ultrasound volumes," *Med Phys,* 2004, 31(4): 744–754.

Sahiner B, Chan HP, Roubidoux MA, Hadjiiski L, Helvie MA, Paramagul C, Bailey J, Nees A, Blane C, "Computer-Aided Diagnosis of Malignant and Benign Breast Masses in 3D Ultrasound Volumes: Effect on Radiologists' Characterization Accuracy," *Radiology (submitted)* 2004.

Conference Abstracts:

Sahiner B, Chan HP, Hadjiiski LM, Roubidoux MA, Paramagul CP, Helvie MA, et al., "The effect of a multi-modality computer classifier on radiologists' accuracy in characterizing breast masses using mammograms and volumetric ultrasound images: An ROC study," to be presented at the *90th Scientific Assembly and Annual Meeting of the Radiological Society of North America*, Chicago, IL, Nov. 28-Dec 3, 2004.

Conference Proceedings:

Sahiner B, Chan HP, Hadjiiski LM, Roubidoux MA, Paramagul C, Helvie MA, Zhou C, "Multi-modality CAD: Combination of computerized classification techniques based on mammograms and 3D ultrasound volumes for improved accuracy in breast mass characterization," *Proc. SPIE Medical Imaging,* 2004, 5370:67-74.

Sahiner B, Chan HP, Hadjiiski LM, Roubidoux MA, Paramagul C, Helvie MA, LeCarpentier GL, "Fusion of mammographic and sonographic computer-extracted features for improved characterization of breast masses," *Digital Mammography IWDM 2004: 7th International Workshop on Digital Mammography*, (in press).

(8) Conclusions

As a result of the support by the USAMRMC BCRP grant, in the third year of this project, we have (1) analyzed the observer performance study conducted in year 2 (radiologists reading 3D US volumes without and with CAD) and submitted a journal paper based on this analysis; (2) evaluated the classification accuracy of mammographic CAD, and compared it to the accuracy of US CAD; (3) identified preferred algorithms and parameters for the design of the computer classifier (multi-modality CAD) that combines the US and mammography information; and (4) performed preliminary studies for the evaluation of the potential improvement when the radiologists use multi-modality CAD.

Our results are very encouraging. We have found that CAD can significantly improve radiologists' characterization accuracy of sonographic breast masses. Although this finding is important, its clinical significance is limited,because radiologists use both mammograms and US images to evaluate masses. We have therefore started to conduct on observer study in which the computer uses both modalities for mass characterization, and radiologists read both modalities with and without the aid of multi-modality CAD. We have found that all five radiologists who completed the study so far showed improvement with CAD. We believe that participation of additional radiologists and further analysis of this study will reveal the potential of multi-modality CAD. Further improvement of the 3D ultrasound characterization methods and improved methods for combination with mammographic computer image analyses can provide radiologists with a powerful aid for decision making, which may help reduce unnecessary biopsies and improve patient care.

(9) References

1. B. Sahiner, H.P. Chan, M.A. Roubidoux, L.M. Hadjiiski, M.A. Helvie, C. Paramagul, J. Bailey, A. Nees, and C. Blane, "Computer-Aided Diagnosis of Malignant and Benign Breast Masses in 3D Ultrasound Volumes: Effect on Radiologists' Characterization Accuracy," Radiology, (submitted), (2004).

2. B. Sahiner, H.P. Chan, M.A. Roubidoux, M.A. Helvie, L.M. Hadjiiski, A. Ramachandran, G.L. LeCarpentier, A. Nees, C. Paramagul, and C.E. Blane, "Computerized characterization of breast masses on 3-D ultrasound volumes," Medical Physics, 31, 744-754 (2004).

3. Y. Jiang, C.E. Metz, and R.M. Nishikawa, "A receiver operating characteristic partial area index for highly sensitive diagnostic tests," Radiology, 201, 745-750 (1996).

4. D.D. Dorfman, K.S. Berbaum, and C.E. Metz, "ROC rating analysis: Generalization to the population of readers and cases with the jackknife method," Investigative Radiology, 27, 723-731 (1992).

5. *American College of Radiology Breast Imaging Reporting and Data System Atlas (BI-RADS Atlas)*. 2003, Reston, VA: American College of Radiology.

6. E.A. Sickles, "Nonpalpable, circumscribed, noncalcified solid breast masses: likelihood of malignancy based on lesion size and age of patient.," Radiology, 192, 439-442 (1994).

7. B. Sahiner, H.P. Chan, L.M. Hadjiiski, M.A. Roubidoux, C. Paramagul, M.A. Helvie, and C. Zhou, "Multi-modality CAD: Combination of computerized classification techniques based on mammograms and 3D ultrasound volumes for improved accuracy in breast mass characterization," Proceedings of the SPIE - Medical Imaging, 5370, 67-74 (2004).

8. B. Sahiner, H.P. Chan, L.M. Hadjiiski, M.A. Roubidoux, C. Paramagul, M.A. Helvie, and G.L. LeCarpentier, "Fusion of mammographic and sonographic computer-extracted features for improved characterization of breast masses," Proceedings of the 7th International Workshop on Digital Mammography, (in press), (2004).

(10) Appendix

Copies of the following publications are enclosed with this report:

(1) Sahiner B, Chan HP, Roubidoux MA, Helvie MA, Hadjiiski LM, Ramachandran A, LeCarpentier GL, Nees A, Paramagul C, Blane CE, "Computerized characterization of breast masses on 3-D ultrasound volumes," *Med Phys,* 2004, 31(4): 744–754.

(2) Sahiner B, Chan HP, Hadjiiski LM, Roubidoux MA, Paramagul C, Helvie MA, Zhou C, "Multi-modality CAD: Combination of computerized classification techniques based on mammograms and 3D ultrasound volumes for improved accuracy in breast mass characterization," *Proc. SPIE Medical Imaging,* 2004, 5370:67-74.

(3) Sahiner B, Chan HP, Roubidoux MA, Hadjiiski L, Helvie MA, Paramagul C, Bailey J, Nees A, Blane C, "Computer-Aided Diagnosis of Malignant and Benign Breast Masses in 3D Ultrasound Volumes: Effect on Radiologists' Characterization Accuracy," *Radiology (submitted)* 2004.

(4) Sahiner B, Chan HP, Hadjiiski LM, Roubidoux MA, Paramagul CP, Helvie MA, et al., "The effect of a multi-modality computer classifier on radiologists' accuracy in characterizing breast masses using mammograms and volumetric ultrasound images: An ROC study," Submitted for presentation at the *90th Scientific Assembly and Annual Meeting of the Radiological Society of North America*, Chicago, IL, Nov. 28-Dec 3, 2004.

(5) Sahiner B, Chan HP, Hadjiiski LM, Roubidoux MA, Paramagul C, Helvie MA, LeCarpentier GL, "Fusion of mammographic and sonographic computer-extracted features for improved characterization of breast masses," to appear in proceeding of *Digital Mammography IWDM 2004: 7th International Workshop on Digital Mammography.*

Computerized characterization of breast masses on three-dimensional ultrasound volumes

Berkman Sahiner,[a)] Heang-Ping Chan, Marilyn A. Roubidoux, Mark A. Helvie,
Lubomir M. Hadjiiski, Aditya Ramachandran, Chintana Paramagul,
Gerald L. LeCarpentier, Alexis Nees, and Caroline Blane
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109-0904

(Received 5 August 2003; revised 17 December 2003; accepted for publication 23 December 2003; published 11 March 2004)

We are developing computer vision techniques for the characterization of breast masses as malignant or benign on radiologic examinations. In this study, we investigated the computerized characterization of breast masses on three-dimensional (3-D) ultrasound (US) volumetric images. We developed 2-D and 3-D active contour models for automated segmentation of the mass volumes. The effect of the initialization method of the active contour on the robustness of the iterative segmentation method was studied by varying the contour used for its initialization. For a given segmentation, texture and morphological features were automatically extracted from the segmented masses and their margins. Stepwise discriminant analysis with the leave-one-out method was used to select effective features for the classification task and to combine these features into a malignancy score. The classification accuracy was evaluated using the area A_z under the receiver operating characteristic (ROC) curve, as well as the partial area index $A_z^{(0.9)}$, defined as the relative area under the ROC curve above a sensitivity threshold of 0.9. For the purpose of comparison with the computer classifier, four experienced breast radiologists provided malignancy ratings for the 3-D US masses. Our dataset consisted of 3-D US volumes of 102 biopsied masses (46 benign, 56 malignant). The classifiers based on 2-D and 3-D segmentation methods achieved test A_z values of 0.87±0.03 and 0.92±0.03, respectively. The difference in the A_z values of the two computer classifiers did not achieve statistical significance. The A_z values of the four radiologists ranged between 0.84 and 0.92. The difference between the computer's A_z value and that of any of the four radiologists did not achieve statistical significance either. However, the computer's $A_z^{(0.9)}$ value was significantly higher than that of three of the four radiologists. Our results indicate that an automated and effective computer classifier can be designed for differentiating malignant and benign breast masses on 3-D US volumes. The accuracy of the classifier designed in this study was similar to that of experienced breast radiologists.
[DOI: 10.1118/1.1649531]

Key words: computer-aided diagnosis, 3-D ultrasound, breast mass characterization, segmentation

I. INTRODUCTION

The importance of early breast cancer detection requires a vigorous approach to the characterization of breast lesions. At present, the positive biopsy rate for nonpalpable breast lesions as well as for nonpalpable breast masses is between 15%–30%.[1–4] This means that 70%–85% of breast biopsies are performed for benign lesions. In order to reduce patient anxiety and morbidity, as well as to decrease health care costs, it is desirable to reduce the number of benign biopsies without missing malignancies. Computer-aided diagnosis (CAD) can provide a consistent and reproducible second opinion to the radiologists, and has a potential to assist them in reducing benign biopsies. Recent studies on the computerized classification of breast masses based on mammographic image features suggest that the radiologists' performance may be significantly improved if they are aided by a well-trained CAD system.[5–7] Breast ultrasound (US) is an important imaging modality for the characterization of breast masses as malignant and benign. An objective and reproducible second opinion from a computer classifier for the classification of breast masses based on US image features may be an important addition to CAD tools being developed for mammographic image analysis.

Breast US is widely accepted as a highly accurate modality for the differentiation of cystic and noncystic masses. As a result of technological improvements and more sophisticated utilization by radiologists, US has been gaining popularity for the characterization of noncystic, or solid, breast masses. By combining several ultrasonic characteristics, Stavros *et al.*[8] achieved a specificity of 98.4% and a sensitivity of 68.7% on a dataset of 750 solid breast masses. Using strict criteria for a benign diagnosis, Skaane *et al.*[9] achieved a positive predictive value of 66% and a negative predictive value of 98% for the differentiation of fibroadenoma and invasive ductal carcinoma on sonograms. Recently, Taylor *et al.* investigated whether the complementary use of US imaging could decrease the biopsy of benign, noncystic masses. On a dataset of 761 biopsied masses, they found that the addition of US evaluation to mammography alone could increase the specificity from 51.4% to 63.8% while slightly increasing the sensitivity from 97.1% to 97.9%.[10] In our study we aim at developing techniques for the computerized characterization of solid breast masses, which may eventually improve the radiologists' accuracy in this difficult and important task.

A number of researchers have recently investigated the application of CAD to breast US images.[11–14] Chen *et al.*[12] extracted autocorrelation features from rectangular regions of interest (ROIs) containing solid breast masses. Using a neural network classifier, they obtained an area A_z under the receiver operating characteristic (ROC) curve of 0.956 for classification of a dataset of 140 biopsy-proven masses as malignant or benign. Horsch *et al.*[13] developed an automated segmentation method for delineating the mass boundaries, and compared its characterization accuracy on different subsets with that obtained from manual segmentation. Using manual and automated segmentation methods, they obtained A_z values of 0.91 and 0.87, respectively, in the task of differentiating all malignant and benign lesions in their dataset, and 0.88 and 0.82, respectively, in the task of differentiating the subset of malignant and benign solid lesions. Chen *et al.*[14] used morphological features extracted from manually segmented mass boundaries for classification. Using a neural network classifier, they obtained an A_z of 0.959 for classification of a dataset of 271 biopsy-proven masses as malignant or benign.

A 3-D US is rapidly gaining popularity as it moves out of the research environment and into the clinical setting.[15] A computerized analysis of 3-D US images may be useful for two reasons. First, 3-D or volumetric US data may be more time consuming for a radiologist to interpret, thus making CAD more desirable. Second, 3-D or volumetric US provides more data and better statistics, which should improve statistical image analysis.

In clinical practice, breast US may be performed in different ways. In many breast imaging clinics, the US examination is performed by a US technologist. Once the technologist locates the mass, and determines the appropriate settings for optimal image quality, representative static US images of the mass are printed on hardcopy film. The radiologist only reads the images chosen by the technologist. A second possibility is that the US scan is videotaped by the technologist and the radiologist reads the examination on a video display. In a third method, a radiologist will perform the US examination interactively and optimize the image quality by changing the probe angle, direction, and US machine settings. Since the US image quality is operator dependent, the way in which the examination is performed may have an impact on the diagnostic accuracy. At our institution, the third method is employed. As described in Sec. II, the data acquisition system in this study did not permit interactive modification during 3-D image acquisition. As a result, the data that was used by the computer and the radiologists for mass characterization in this study may not be as informative as the data that the radiologists could have obtained by examining the patient interactively. However, since the mass is entirely imaged in the 3-D dataset, our data should be at least comparable to that obtained by using the first method described above.

In this study, we investigated the computerized characterization of noncystic breast masses as malignant and benign in 3-D US images. We developed a 3-D segmentation method to delineate the masses. Morphological and texture features were extracted from the mass and its margins for classification. A linear classifier was used to merge the features into a malignancy score. The classification accuracy was evaluated by ROC methodology. The ROC curves of the computer and four experienced breast radiologists were compared. To our knowledge, this is the first study on 3-D US images that investigates a computer segmentation method followed by a computer classifier for breast cancer characterization.

II. METHODS

A. Dataset

Institutional review board approval was obtained prior to the commencement of this investigation. The images used in this study were acquired between 1998 and 2002. Our study group was 102 women (average age: 51 years) who had a solid mass deemed suspicious or highly suggestive of malignancy. All patients underwent biopsy or fine needle aspiration. Fifty-six masses were malignant and 46 were benign. Forty-three of the malignancies were invasive ductal carcinoma, five were invasive lobular carcinoma, one was medullary carcinoma, three were ductal carcinoma *in-situ*, and four were other invasive carcinoma. Of the benign masses, the majority were fibroadenoma ($N=18$) and fibrocystic disease ($N=11$). The mean equivalent lesion diameter was 1.28 cm (standard deviation=0.78 cm).

The 3-D US data were acquired using an experimental system that was previously developed and tested at our institution.[16,17] The 3-D system consisted of a commercially available US scanner (GE Logiq 700 with an M12 linear array transducer), a mechanical transducer guiding system, and a computer workstation. The linear array transducer was operated at 11 MHz. The technologist was free to set the focal distance and the overall gain adjustment to obtain the best possible image. Before 3-D image acquisition, the technologist used clinical US and mammogram images to identify the suspicious mass. During 3-D image acquisition, the technologist manually translated the transducer linearly in the cross-plane, or the z direction, while the image acquisition system recorded 2-D B-mode images in the image scan plane (x-y plane). The 2-D images were obtained at approximately 0.5 mm incremental translations, which were measured and recorded using a translation sensor. The number of 2-D slices was typically around 90, and varied depending on the lesion size. The maximum distance between two 2-D slices was 0.5 mm, and some of the distances were slightly less than 0.5 mm. The scanned breast region measured typically 4.5 cm long by 4.0 cm wide by 4.0 cm deep. The typical pixel size in a slice was approximately 0.11 mm.

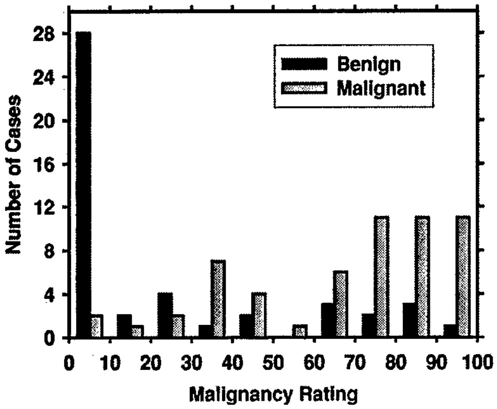


FIG. 1. The distribution of the malignancy rating of the masses in our dataset based on the appearance on US images, by an experienced radiologist. 1: Very likely benign; 100: very likely malignant.

The B-mode images were recorded into a buffer in the US scanner. After data acquisition, the images and the position data were transferred digitally to a workstation, where individual planes were cropped and stacked to form a 3-D volume. The biopsied mass in each volume was identified by a MQSA (Mammography Quality Standards Act) qualified radiologist (RAD1) using clinical US and mammographic images to confirm that the 3-D images contained the suspicious mass. The likelihood of malignancy for each mass, based on the 3-D US image alone, was rated by the same radiologist on a scale of 1 to 100, where a higher number corresponded to a higher likelihood of malignancy. The distribution of the ratings for the malignant and benign masses is shown in Fig. 1. The radiologist was also asked to fit a 3-D ellipsoid to the mass. The 3-D ellipsoid was used to initialize the computerized mass segmentation described in the next section. The best fit was obtained by scaling, rotating, and translating an ellipsoid superimposed on the 3-D dataset using a dynamic object manipulation tool developed for this purpose.

B. Mass segmentation

We investigated the use of 2-D and 3-D active contour models for the segmentation of mass boundaries.[18] An active contour model is a high-level segmentation method that uses energy terms derived from the image gray-level information as well as the *a-priori* knowledge about the object to be segmented for accurate segmentation. The segmentation problem is defined as an energy minimization problem. In order for the model to lock onto the contours in the image, the image-based energy terms, also referred to as the external energy terms, are usually defined in terms of the image gray levels and the image gradient magnitude. The *a-priori* knowledge of the object shape is used to define internal energy terms related to features such as the continuity and the smoothness of the contour to constrain the segmentation problem. These terms can compensate for noise or apparent gaps in the image gradients, which often mislead segmentation methods that do not use *a-priori* information.

In a 2-D segmentation problem, the contour of the object can be represented by V vertices, (i_ν,j_ν), $\nu=1,...,V$, where i and j represent the two dimensions of the image. In the discrete formulation of the active contour model, the total energy to be minimized is defined as

$$E=\sum_{\nu=1}^{V} E(\nu), \tag{1}$$

where $E(\nu)$ is the energy at vertex (i_ν,j_ν). $E(\nu)$ is defined as the sum of the internal and external energy terms,

$$E(\nu)=\sum_{m=1}^{M} w_m E_m(\nu), \tag{2}$$

where $E_m(\nu)$ is the mth energy term at vertex ν, and w_m is the weight of the mth energy term. In our 2-D active contour model, we used four internal and external energy terms ($M=4$). The energy terms E_1, E_2, E_3, and E_4 were determined by the gradient magnitude of the image and the continuity, smoothness, and balloon energy of the contour, respectively.

To obtain the image gradient magnitude, the image $A(i,j)$ was first filtered using a Gaussian smoothing filter,

$$H(i,j)=e^{-(i^2+j^2)/2\sigma^2}, \tag{3}$$

where $\sigma^2=6$. The resulting filtered image $B(i,j)$ was further processed using Sobel filters $S_x(i,j)$ and $S_y(i,j)$, defined as

$$S_x=\begin{bmatrix}-1 & 0 & 1\\ -2 & 0 & 2\\ -1 & 0 & 1\end{bmatrix} \quad \text{and } S_y=\begin{bmatrix}-1 & -2 & -1\\ 0 & 0 & 0\\ 1 & 2 & 1\end{bmatrix}, \tag{4}$$

which calculated the x- and y-direction gradients, $G_x(i,j)$ and $G_y(i,j)$, respectively. The image gradient magnitude at vertex $\nu=(i_\nu,j_\nu)$ was computed as

$$E_1(\nu)=\sqrt{G_x(i_\nu,j_\nu)+G_y(i_\nu,j_\nu)}. \tag{5}$$

The weight of the gradient energy was defined to be a negative number; thus, minimizing w_1E_1 attracted the contour to image edges.

To find the continuity energy term, we first computed the average line segment length $\bar{d}$ as

$$\bar{d}=\frac{\Sigma_{\nu=1}^{V} d(\nu)}{V}, \tag{6}$$

where

$$d(\nu)=\begin{cases}\sqrt{(i_\nu-i_{\nu+1})^2+(j_\nu-j_{\nu+1})^2}, & \nu=1,2,...,V-1,\\ \sqrt{(i_\nu-i_0)^2+(j_\nu-j_0)^2}, & \nu=V.\end{cases} \tag{7}$$

The continuity energy term was defined as

$$E_2(\nu)=|d(\nu)-\bar{d}|. \tag{8}$$

Minimizing the continuity energy helped the vertices maintain regular spacing along the contour.

The curvature term, $E_3(\nu)$, was approximated by the second derivative of the contour,

$$E_3(\nu)=\sqrt{(i_{\nu-1}-2i_\nu+i_{\nu+1})^2+(j_{\nu-1}-2j_\nu+j_{\nu+1})^2}. \quad (9)$$

When the vertices were spaced regularly along the contour, this term would be large when the angle at vertex ν was small.[19] By discouraging small angles at vertices, this term attempted to smooth the contour.

The balloon energy $E_4(\nu)$ pushed the contour outward or pulled it inward, depending on whether w_4 was positive or negative, respectively, along a path normal to the contour. This energy term helped the active contour traverse spurious, isolated, or weak image edges, and countered its tendency to shrink. The resulting model was reported to be more robust to the initial position and image noise.[20]

To solve the energy minimization problem, we have chosen the iterative method proposed by Williams and Shah.[19] The contour is first initialized by defining V vertices (i_ν,j_ν), $\nu=1,...,V$. At a given iteration, the method visits each vertex (i_ν,j_ν). Let $\mathbf{D}(\nu)$ represent the set of pixels (i',j') in a $(2M+1)x(2M+1)$ neighborhood centered around (i_ν,j_ν). For each pixel in $\mathbf{D}(\nu)$, the sum $\Sigma_m w_m E_m$ is computed, and the vertex (i_ν,j_ν) is moved to the (i'^*,j'^*) location that minimizes this sum. The definitions of the energy terms E_1, E_2, and E_3 are given above. The balloon energy E_4 was defined as $E_4=\cos\theta$, where θ represents the angle between the normal vector to the curve at vertex ν and the vector $(i'-i_\nu,j'-j_\nu)$. After the minimization is performed locally at vertex (i_ν,j_ν), the algorithm moves to the vertex $(i_{\nu+1},j_{\nu+1})$. The method converges when no vertex changes location at a given iteration. In practical implementation, iterations may be stopped when a large, predetermined percentage of vertices stop moving. The cross section of the radiologist-defined ellipsoid with each image slice was used for initializing the contour.

When the 2-D active contour model described above is applied to a 3-D dataset, segmentation is performed independently on each slice of the 3-D volume. However, this kind of segmentation ignores the continuity of the object across slices. When the slice spacing is small compared to the rate of change of the object shape, it is expected that the shape of the object is unlikely to change drastically from one slice to the next. Our 3-D active contour model is aimed at using the shape information across the 3-D slices to improve upon the 2-D active contour model. Our 3-D active contour model was defined by including in the curvature energy term, an additional component related to the smoothness of the mass in the z direction. Let $(i_{\nu,k},j_{\nu,k})$ denote the νth vertex in image slice k. The curvature energy in our 3-D active contour model was defined as

$$\begin{aligned}E_3(\nu)&\\ &=\sqrt{(i_{\nu-1,k}-2i_{\nu,k}+i_{\nu+1,k})^2+(j_{\nu-1,k}-2j_{\nu,k}+j_{\nu+1,k})^2}\\ &\quad+\alpha\sqrt{(i_{\nu,k-1}-2i_{\nu,k}+i_{\nu,k+1})^2+(j_{\nu,k-1}-2j_{\nu,k}+j_{\nu,k+1})^2},\end{aligned} \quad (10)$$

where α was the weight of the out-of-plane component of the curvature relative to the in-plane component. The out-of-plane component forced the contour to be smooth in the z direction. Our implementation of the 3-D active contour model started by optimizing the contour in the first slice of the 3-D dataset $(k=1)$. Since slice $k=0$ did not exist, we assumed that $(i_\nu,j_\nu,0)=(i_\nu,j_\nu,1)$ for all ν. The contour optimization in slice $k=1$ followed the steps described above for 2-D active contours, except that the curvature energy was replaced by Eq. (10). After the contour was optimized for slice $k=1$, the optimization was performed for slice $k=2$, and so on, until the contours were optimized for all slices. This constituted one 3-D iteration. The 3-D model repeated the 3-D iterations until there was no movement of the vertices for the 3-D contour, or when a predetermined percentage of vertices stopped moving. Similar to our 2-D active contour, the 3-D active contour was initialized using the radiologist-defined ellipsoid.

We did not employ an optimization method for determining the active contour weights because automatic optimization required the comparison of the automated contour with a gold standard such as the radiologist's manual segmentation for training. The "true" borders of many masses on US images were not well defined, even to experienced radiologists. Furthermore, the features that we designed did not require a border that followed the detailed boundary of an ill-defined or a spiculated mass. We therefore used more subjective judgment on the "goodness of segmentation" for the mass boundary based on our experience with the need of the features. To determine the weights for the 2-D model, we started with weights we had previously used for the segmentation of masses on mammograms.[21] We experimentally modified the weights and observed the effect on the segmentation quality for the first 15 volumes in our dataset. We found that the combination $w_1=-1.5$, $w_2=1$, $w_3=2.6$, and $w_4=0.2$ provided a good balance between the smoothness of the contour and its the attraction to the mass borders. These weights were then used for the 2-D segmentation of the entire dataset. For the 3-D active contour model, we maintained the weights at the values that we determined for the 2-D active contour model, and selected $\alpha=0.5$. The choice of α was again based on a qualitative assessment of segmentation on the first 15 cases.

C. Feature extraction

We have evaluated a number of morphological and texture features for characterization of the masses as malignant or benign. Each of the features described below was extracted from every slice where the mass was segmented using either the 2-D or the 3-D automated segmentation algorithm. The features extracted from different slices of the same mass were then combined to define the feature measures (such as mean or maximum) for that mass.

1. Extraction of morphological features

The taller-than-wide shape of a sonographic mass is a good indication of malignancy.[8] This characteristic was defined by the ratio of the widest cross section (W) of the

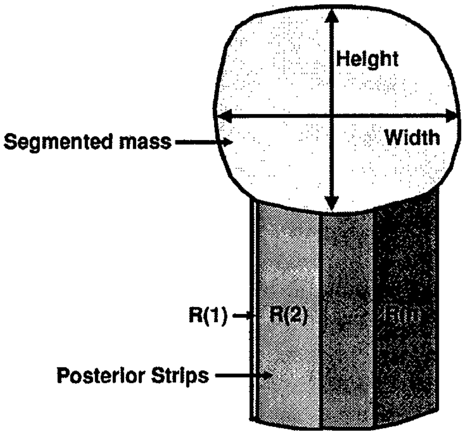


FIG. 2. The definition of the width-to-height and PSF features. The width-to-height feature was defined as the ratio of the widest cross section of the segmented mass shape in the image plane to the tallest cross section. The PSF feature was defined by first finding the average gray value in the posterior strips $\overline{R(i)}$, $i=1,\ldots,n$, then finding the minimum of $\overline{R(i)}$ among the n strips, and finally by normalizing this value by the average gray value within the segmented mass.

automatically segmented lesion shape to the tallest cross section (T) in a slice (Fig. 2). Another feature that has been reported to be useful for differentiation of malignant and benign masses is posterior shadowing. In order to define a posterior shadowing feature (PSF), we first calculated the mean pixel value $\overline{R(i)}$ in overlapping vertical strips $R(i)$, $i=1,\ldots,n$ posterior to the mass, as shown in Fig. 2. The width W_R of a strip was equal to one-fourth of the width of the mass ($W/4$), and the height of the strip was equal to the height of the mass (T). The left and right edges of strips $R(i)$ and $R(i+1)$ differed by one pixel. In other words, the strip $R(i+1)$ was obtained by moving the strip $R(i)$ to the right by one pixel, while, of course, the strip remained posterior to the mass and its height remained as T. In order to exclude the bilateral posterior shadowing artifacts that are sometimes associated with fibroadenomas, the strips were defined only posterior to the central $3W/4$ portion of the mass (Fig. 2). The minimum value of these averages, $\min\{\overline{R(i)},i=1,\ldots,n\}$, was the darkest posterior strip. The PSF was defined as the normalized average gray-level difference between the interior of the segmented mass and the darkest posterior strip,

$$\mathrm{PSF}=\frac{\bar{M}-\min\{\overline{R(i)},i=1,\ldots,n\}}{\bar{M}}, \tag{11}$$

where $\bar{M}$ denotes the mean gray level value inside the segmented mass.

2. Extraction of texture features

The features used in this study were extracted from spatial gray-level dependence (SGLD) matrices, or co-occurrence matrices, derived from 2-D slices of the 3-D dataset. The (i,j)th element of the co-occurrence matrix is the relative frequency with which two pixels: one with gray level i and the other with gray level j, separated by a pixel pair distance d in a direction θ occur in the image. Features extracted from SGLD matrices of US images have been shown to be useful in the classification of malignant and benign breast masses on mammograms in previous studies.[22] In this study, six texture feature measures that are invariant under linear, invertible gray scale transformations were extracted. These features were information measures of correlations 1 and 2 (IMC1 and IMC2), difference entropy (DFE), entropy (ENT), energy (ENE), and sum entropy (SME). The mathematical definitions of these features can be found in the literature.[23] Although many gray scale transformations may not be invertible due to pixel saturation or roundoff, these features are largely independent of the gray-level gain adjustments.

It is known that the margin characteristics of a mass are very important for its characterization, and previous studies have indicated that texture features extracted from the mass margins are effective for classification.[24] For this reason, the texture features in this study were extracted from two disk-shaped regions containing the boundary of each mass, as well as presumably mass and normal tissue adjacent to the boundary of the mass. These regions followed the contour determined by the active contour model, as shown in Fig. 3. The areas for the upper and lower disk-shaped regions were chosen to be equal, and their sum was equal to the area of the segmented mass. The pixel pair distances used for SGLD matrix computation were chosen to be $d=2$, 4, and 6. Two pixel pair angles, $\theta=0°$ and $\theta=90°$, were evaluated for each d in both regions. The number of SGLD matrices computed for a disk-shaped region was therefore 6, and the number of features extracted from an image containing the segmented mass was 72 (6 features, extracted from 6 SGLD matrices in the upper disk-shaped region and the lower disk-shaped region).

D. Classification

The features extracted from different slices of the same mass were combined to define the feature measures for that mass. For the width-to-height feature and the PSF, we computed the mean, variance, minimum, and maximum of the extracted value from each slice containing the mass. Therefore eight morphological feature measures were defined for each mass. For texture features, we only computed the mean, hence 72 texture feature measures were defined for each mass.

Fisher's linear discriminant analysis (LDA)[25] was used for combining the features into a discriminant score. Since the number of available features in the feature space was relatively high compared with the number of available cases, stepwise feature selection[26] was used in order to reduce the number of the features and to obtain the best feature subset to design an effective classifier. For partitioning the dataset into trainers and testers, we used the leave-one-case-out resampling method. Feature selection is performed as part of the classifier design such that both the feature selection and the classifier coefficient estimation procedures were repeated 102 times, as each case was left out once as the test sample. The test discriminant scores were analyzed using ROC

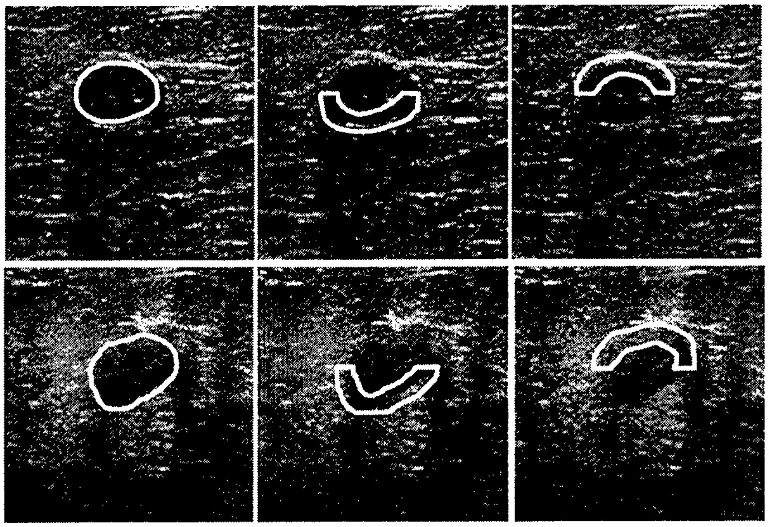


FIG. 3. Left column: The segmented object for a malignant mass (upper row) and a benign mass (lower row). Middle and right columns: The lower and upper disk-shaped regions from which texture features were extracted.

methodology.[27] The classification accuracy was evaluated using the area under the ROC curve, A_z, as well as the partial area index, $A_z^{(0.9)}$. $A_z^{(0.9)}$ is defined as the area under the ROC curve above a sensitivity threshold of 0.9 ($\mathrm{TPF}_0=0.9$) normalized to the total area above TPF_0, which is equal to $(1-\mathrm{TPF}_0)$.[28]

E. Malignancy ranking by radiologists

Although all the cases in our dataset were suspicious enough to warrant biopsy or fine needle aspiration, the degree of difficulty of our cases can best be measured by investigating the accuracy of the radiologists in classifying the cases in our dataset as malignant or benign. As described in Sec. II B, one radiologist (RAD1) who was familiar with the clinically obtained images had initially provided a malignancy rating. To compare with the computer's accuracy, we are interested in measuring the accuracy of other radiologists, who would not be biased by memory or familiarity with the cases. For this purpose, we have developed an interactive graphical user interface with which the radiologists could navigate through 3-D volumes, adjust the window and level of the displayed images, and enter a malignancy rating between 1 and 100 (a higher rating indicating a higher likelihood of malignancy) when they finish examining a case. Three additional radiologists (RAD2–RAD4) participated in the malignancy rating study. The radiologists RAD1–RAD4

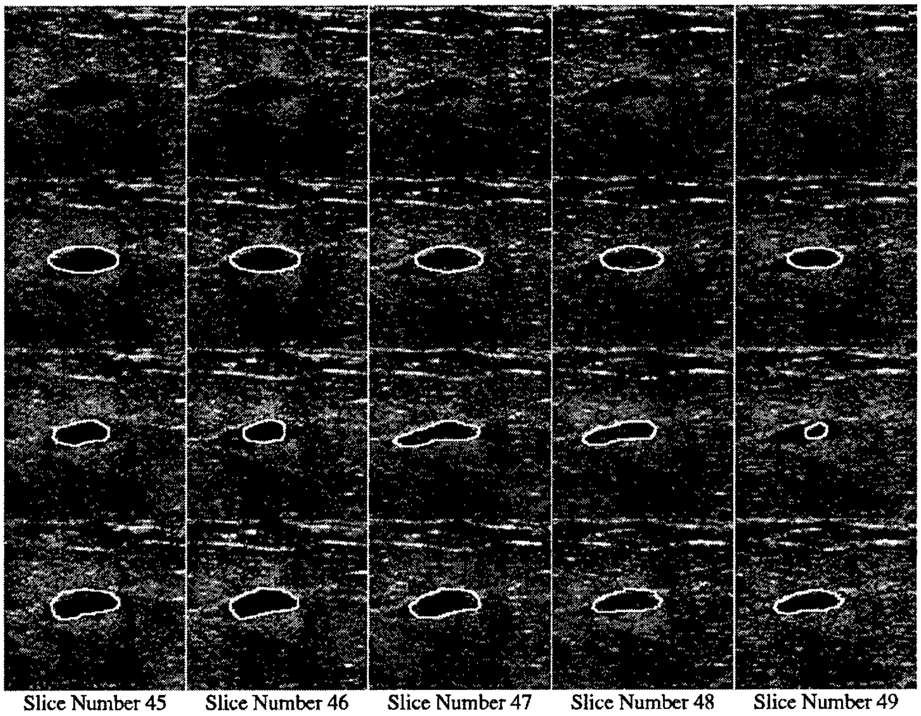


FIG. 4. Row 1: Five original slices of a breast mass that was visible on a total of ten US slices; row 2: The cross section of the initial 3-D ellipsoid at each slice; row 3: The result of the 2-D active contour segmentation method; row 4: The result of the 3-D active contour segmentation method. Note that the 2-D segmentation method missed part of the mass on slice 46. The 3-D segmentation method, apparently using the information from slices 45 and 47, was able to provide better segmentation on slice 46.

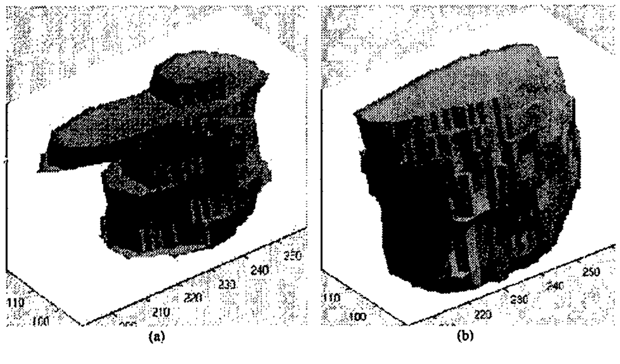


FIG. 5. 3-D rendering of the segmented object for the mass shown in Fig. 4. (a) 2-D active contour segmentation; (b) 3-D active contour segmentation.

were either fellowship trained in breast imaging or had over 25 years of experience in breast imaging. All four radiologists were MQSA qualified and their experience in mammographic and US interpretation ranged from 2 to 25 years (mean, 11.3 years). The location of the mass center, as determined by RAD1, was displayed on each slice, so that all the radiologists would rank the same mass if more than one mass existed in the volume. There was no time limitation for the radiologists to read a case. The case reading order was randomized for each radiologist. The malignancy rating was entered by means of a slide bar. Before participating in the study, the radiologists were trained on five cases that were not part of the test dataset described in Sec. II A. The malignancy rating study was intended to measure the difficulty of the dataset, and was not intended to measure how the radiologists' interpretation would be affected by CAD. Therefore, the computer classification results were not displayed to the radiologists in this study.

III. RESULTS

We evaluated the accuracy of characterization based on both 2-D and 3-D active contour segmentation methods. Rows 1 to 4 of Fig. 4 show the original images, radiologist-defined ellipsoid, 2-D active contour results, and 3-D active contour results for five consecutive slices of a mass that was visible on a total of 10 slices. Figure 5 shows a 3-D rendering of the segmented object using the 2-D and 3-D active contour models. It is seen from Fig. 5 that the shape of the object segmented by the 3-D active contour model is smoother in the z direction.

Table I shows the range (minimum and maximum) of the A_z values provided by each texture feature alone, extracted from the upper and lower disk-shaped regions determined by the 2-D and 3-D active contour models. The ranges in this table are for different pixel pair distances and directions used in extracting the same feature (e.g., IMC1). Table II shows the range of A_z values provided by each morphological feature alone, using the 2-D and 3-D active contour models. The ranges in Table II are for different methods of combining the features extracted from individual slices, i.e., mean, variance, minimum, and maximum. The most discriminatory feature in this study was the IMC1 feature ($d=6$, $\theta=0°$, extracted from the upper disk-shaped region segmented by the 3-D method) with an A_z value of 0.76.

TABLE I. The range of A_z values for different texture features extracted from the lower and upper disk-shaped regions using the 3-D and 2-D segmentation methods. For each particular texture feature (e.g., IMC1 feature at pixel-pair distance $d=2$, and direction $\theta=0°$), the feature values from all the slices containing the segmented mass were averaged before computing the A_z value. The range indicates the minimum–maximum A_z values for a particular feature among the parameters $d=2$, 4, 6 and $\theta=0°$, 90°.

	3-D segmentation		2-D segmentation	
Texture feature	Upper	Lower	Upper	Lower
IMC1	0.66–0.76	0.58–0.67	0.65–0.72	0.59–0.66
IMC2	0.65–0.75	0.58–0.65	0.65–0.73	0.61–0.67
DFE	0.58–0.68	0.61–0.67	0.56–0.68	0.62–0.70
ENT	0.59–0.64	0.55–0.60	0.62–0.69	0.58–0.62
ENE	0.57–0.63	0.53–0.60	0.53–0.60	0.50–0.54
SME	0.52–0.58	0.51–0.56	0.57–0.64	0.52–0.57

TABLE II. The range of A_z values for the width-to-height feature and posterior shadowing feature (PSF) extracted using the 3-D and 2-D segmentation methods. The range indicates the minimum–maximum A_z values among the mean, variance, minimum, and maximum of each feature extracted from each slice containing the segmented mass.

Morphological feature	3-D segmentation	2-D segmentation
Width-to-height	0.58–0.73	0.54–0.69
PSF	0.53–0.66	0.53–0.59

When stepwise LDA was used to combine the features into a discriminant score in the 102 leave-one-case-out training subsets, an average of 6.09 and 7.98 features were selected with the 2-D and 3-D segmentation methods, respectively. For the 2-D segmentation method, the most frequently selected features were two IMC1 features, two IMC2 features, one DFE feature, and one width-to-height feature. For the 3-D segmentation method, the most frequently selected features were two IMC1 features, two IMC2 features, one DFE feature, one ENT feature, one PSF feature, and one

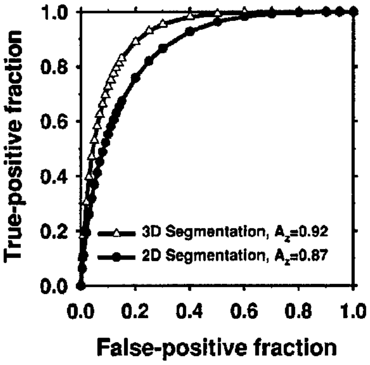


FIG. 6. The test ROC curves obtained by the classifiers that were based on features extracted from the 2-D ($A_z=0.87$) and 3-D ($A_z=0.92$) active contour models. The difference between the two A_z values did not achieve statistical significance ($p=0.07$).

TABLE III. The dependence of the computer classification accuracy on the variation of the initial contour. The effects of three transformation parameters, namely, scaling, translation, and rotation of the initial ellipsoid, was investigated by moving the initial ellipsoid using one of these three parameters at a time. A translation by ±10 pixels in the image plane corresponded to approximately ±1 mm.

Scale	Rotation (degrees)	x-translation (pixels)	y-translation (pixels)	A_z
1	0	0	0	0.92±0.03
1.3	0	0	0	0.89±0.03
0.8	0	0	0	0.89±0.03
1	0	10	10	0.90±0.03
1	0	10	−10	0.87±0.04
1	0	−10	10	0.87±0.04
1	0	−10	−10	0.88±0.03
1	15	0	0	0.93±0.02

width-to-height feature. Figure 6 shows the test ROC curves obtained by the LDA using leave-one-case-out resampling for the 2-D and 3-D segmentation methods. The test A_z values for the 2-D and 3-D methods were 0.87±0.03 and 0.92 ±0.03, respectively, and the $A_z^{(0.9)}$ values were 0.51±0.08 and 0.67±0.08, respectively. The difference between the two test A_z values did not achieve statistical significance ($p = 0.07$). Figure 7 shows the distribution of the discriminant scores obtained from the 3-D method for the malignant and benign cases.

In order to investigate the dependence of the classification accuracy on the initialization of the 3-D active contour model, we scaled, rotated, and translated the initial 3-D ellipsoid and repeated the steps of active contour segmentation, feature extraction, and classification for these modified initial ellipsoids. The classification accuracies for these experiments are presented in Table III. None of the differences between the A_z values on Table III achieved statistical significance.

The ROC curves for the radiologists' malignancy ratings are shown in Fig. 8. The computer and radiologist A_z values and $A_z^{(0.9)}$ values are compared in Table IV. The area A_z under the ROC curve for radiologists RAD1–RAD4 varied between 0.84±0.04 and 0.92±0.03, which are lower than or equal to that of the 3-D computer classifier. The average A_z value, obtained by averaging the slope and intercept parameters (a and b in a ROC analysis) of the individual ROC curves was 0.87. The difference between the A_z values of the individual radiologists and the computer classifiers (2-D and 3-D methods) did not reach statistical significance ($p > 0.05$). The $A_z^{(0.9)}$ values of the computer classifiers based on 2-D and 3-D segmentation were consistently higher than those of all four radiologists. The difference between the $A_z^{(0.9)}$ values of only one of the radiologists (RAD4) and the classifier based on 2-D segmentation achieved statistical significance ($p = 0.05$). The differences between the $A_z^{(0.9)}$ values of three of the four radiologists and that of the classifier based on 3-D segmentation were statistically significant ($p = 0.03$, 0.02, and 0.001 for RAD1, RAD2, and RAD4, respectively).

TABLE IV. The area under the ROC curve (A_z), and the area under the ROC curve above a sensitivity threshold of 0.9 ($A_z^{(0.9)}$) for the computer classifier using the 2-D and 3-D active contour segmentation results, and the four radiologists. The radiologists' results that are significantly ($p < 0.05$) different from the 3-D computer results are noted with an asterisk.

	A_z	$A_z^{(0.9)}$
Computer classifier, 2-D segmentation	0.87±0.03	0.51±0.09
Computer classifier, 3-D segmentation	0.92±0.03	0.67±0.08
RAD1	0.85±0.04	0.47±0.10*
RAD2	0.87±0.03	0.38±0.11*
RAD3	0.92±0.03	0.45±0.15
RAD4	0.84±0.04	0.28±0.11*

IV. DISCUSSION

The computer classifier designed in this study to characterize breast masses on US volumes was able to discriminate between malignant and benign masses that were suspicious enough to warrant a biopsy. From Fig. 7, it is observed that if an appropriate decision threshold was chosen for the discriminant scores of the classifier based on 3-D segmentation, more than 43% (20/46) of biopsied benign masses could be correctly identified while no malignant masses were misclassified (at 100% sensitivity). Based on 2-D segmentation, the corresponding percentage of correctly identified benign masses was 35% (16/46).

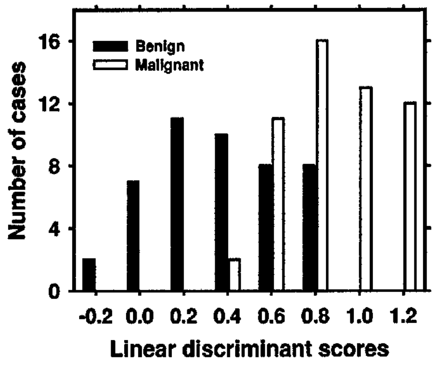


FIG. 7. The distribution of the test discriminant scores for the classifier that was based on 3-D active contour segmentation. By choosing an appropriate decision threshold on these scores (e.g., decision threshold=0.3) more than 43% (20/46) of biopsied benign masses could be correctly identified while no malignant masses would be misclassified.

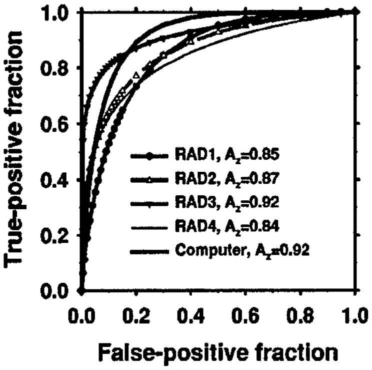


FIG. 8. ROC curves for the computer and for the four radiologists who participated in the malignancy rating experiment. The difference between the computer's A_z value and that of any of the four radiologists did not achieve statistical significance. However, the computer classifier had significantly higher ($p<0.05$) partial area index, $A_z^{(0.9)}$, than three of the four radiologists at high sensitivity (TPF>0.9).

Lesion segmentation is an important task in computerized lesion characterization. The segmentation of US images can be challenging because boundaries are not always conspicuous, due to the noise and contrast characteristics, and the speckled nature of US images. For breast US, an additional source of difficulty is the presence of posterior shadowing artifacts, a major source of which is the US attenuation due to the fibrous stroma caused by the tumor.[29] Previous research on the segmentation of breast masses on US images includes work by Horsch *et al.*,[30] Xiao *et al.*,[31] and Madabhushi *et al.*[32] Their segmentation methods were applied to 2-D US images. In our study, we compared the classification accuracy when 2-D and 3-D active contour models were used for segmentation. The 2-D model provided reasonable segmentation results for many of the masses. However, the 2-D model does not take advantage of the image information in adjacent slices when a particular slice is being segmented. If the 2-D active contour is misled on one slice, there is no interaction from adjacent slices to improve the segmentation. This is illustrated in Fig. 4, row 3. It can be observed that the 2-D segmentation results on slices #45 and #47 are reasonable; however, part of the lesion is missed by the 2-D active contour model on slice #46. Our 3-D active contour model uses the smoothness of the segmented shape in the out-of-plane direction as an interaction term between adjacent slices. The 3-D segmentation results, shown in row 4, are more consistent across slices. Figure 5 compares the segmented object using the 2-D and 3-D methods for the entire lesion, which was visible on a total of ten slices. It is again observed that the lesion shape in the out-of-plane direction is smoother for the 3-D method. Although our classification accuracy using the 3-D method was satisfactory, further improvement may be required for applications such as accurate lesion volume measurement. More sophisticated and inherently 3-D methods, such as deformable surfaces[33] and level set methods, may be good candidates for further improvement.

The texture features in this study were extracted from disk-shaped regions at the upper and lower margins of the mass on each slice. The total area of the two disk-shaped regions was equal to the area of the segmented mass. From Table I, it is observed that a texture feature extracted from the upper disk-shaped region tended to be more discriminatory than the same feature extracted from the lower disk-shaped region. The maximum of the range of A_z values (the second number in each cell) was larger for the upper region in 11 of the 12 comparisons that can be made (6 texture features and 2 segmentation methods). The lower boundaries of many masses were difficult to perceive and hence difficult to automatically segment because of posterior shadowing. This may have contributed to the difference of discrimination ability between the features extracted from the upper and lower regions. Another possible factor may be the changes in the spatial and gray level resolutions in different regions of the US image as the distance from the US probe increases. Further work is underway to investigate the reasons for the apparent lower discrimination ability of the features extracted from the lower disk-shaped regions.

Although the disk-shaped region depends on mass segmentation, there can be a large overlap between the regions from the 2-D and 3-D segmentation results if the objects segmented by the two methods are not very different. From Table I, it can be observed that the ranges of A_z values for 2-D and 3-D segmentation for each texture measure have a large overlap. As mentioned in Sec. III, when the stepwise feature selection method was used for classifier design from 2-D segmentation results, an average of 6.09 features were selected, where the average was computed over the 102 cycles of the leave-one-out partitioning of the dataset. Out of the six most frequently selected features, five were texture features and one was a morphological feature. The IMC1 feature was selected twice (at $d=2$, $\theta=0°$ and $d=6$, $\theta=90°$), the IMC2 feature was selected twice (at $d=2$, $\theta=0°$ and $d=6$, $\theta=0°$), and the DFE feature was selected once (at $d=6$, $\theta=0°$). For 3-D segmentation, out of the eight most frequently selected features, six were texture features, and two were morphological features. The IMC1 feature was selected twice (at $d=2$, $\theta=90°$ and $d=4$, $\theta=0°$), the IMC2 feature was selected twice (at $d=2$, $\theta=0°$ and $d=6$, $\theta=0°$), and the DFE feature was selected once (at $d=6$, $\theta=0°$). Thus, out of 11 most frequently selected texture features (5 for 2-D and 6 for 3-D segmentation), 10 were IMC1, IMC2, or DFE features. The classification accuracy with the stepwise LDA for the 3-D segmentation ($A_z=0.92$) was better than that for 2-D segmentation ($A_z=0.87$). However, the difference did not achieve statistical significance (a two-tailed p value=0.07).

The active contour method requires an initial boundary to start iterating toward the optimal contour. In this study, the initial boundary was defined by a 3-D ellipsoid that approximated the mass shape. The ellipsoid was placed in the volume by one of the radiologists (RAD1) using an interactive graphical user interface (GUI). The radiologist thus had to shift and scale a single object to define the initial contour. Although the error between the true and approximated shapes can be large when a single object is used for approximating the mass, this method was faster than other possible methods that would require initialization on each slice separately, and was therefore preferred. The robustness of the 3-D segmentation method to active contour initialization was studied by translating, rotating, and scaling the 3-D ellipsoid. There are many possibilities as to how these three operations (moving, rotating, and scaling) can be combined to modify the initial ellipsoid. In Table III, the classification results are presented when these three operations are performed one at a time. Row 1 shows the A_z value when the original ellipsoid is used. The ellipsoid was scaled in rows 2–3, translated in rows 4–6, and rotated in row 7. For the magnitudes of scaling, translation, and rotation studied in Table III, the variation of the A_z value was within two standard deviations of the A_z value provided by the LABROC program.[27] In a step toward automating the initialization of the contour, we are currently investigating methods for automatically determining an initial contour from a rectangular box containing the mass.

The comparison of the ROC curves by the radiologists and the computer indicated that the computer can be as effective as the radiologists in differentiating malignant and benign breast masses in this dataset. In fact, the accuracy of the computer classifier using 3-D segmentation was greater than three and equal to one of the radiologists, although the difference between the computer and the individual radiologists in terms of A_z did not achieve statistical significance. Furthermore, from Fig. 8, it is observed that the computer has a tendency to be better at high sensitivity. This was also confirmed by the statistically significant difference between the computer classifier (3-D segmentation method) and three out of the four radiologists when the comparison was based on the $A_z^{(0.9)}$ values. It should be noted that the purpose of our study was not to evaluate our US mass characterization method in a clinical setting. As noted in Secs. I and II, the semiautomated 3-D data acquisition system used in this study is still under investigation and is different from that in current clinical practice. The first difference is that, in our department, radiologists interactively perform handheld US examination themselves, which may yield better image quality and may result in higher characterization accuracy. The second difference is that our study concentrated only on mass characterization of lesions already detected, whereas the actual detection of suspicious masses by US is a very important step in a clinical examination. These other aspects of comparing 3-D US images to US images acquired with current clinical methods are subjects of future investigations.

In this study, the features were extracted from individual US slices and then combined into object-based features, as explained in Sec. II D. Although this method is found to provide effective features in this study, it may not have fully utilized the information available in the 3-D dataset. The potential improvement in classification accuracy by using truly 3-D features, for example, texture features extracted from 3-D SGLD matrices, needs to be investigated. Furthermore, in clinical practice, the decision about whether the mass is malignant or benign is made using both mammographic and US image information, as well as other pertinent patient information. A study is currently underway in our laboratory to design a classifier that combines computer-extracted features or scores from these two imaging modalities.

V. CONCLUSION

A computer segmentation and classification method has been developed for the task of the characterization of breast masses on 3-D US images. On a dataset of 102 biopsy-proven masses the classifier achieved an A_z value of 0.92. The average A_z value of four experienced radiologists on the same data set was 0.87. The computer classifier was more accurate than three and equal to one of the four radiologists participated in the study. However, the difference between the A_z values of the computer and the individual radiologists did not achieve statistical significance for this dataset. At high sensitivity, the computer classifier was consistently more accurate than all four radiologists and achieved statistical significance ($p<0.05$) for the difference in $A_z^{(0.9)}$ from three of the four radiologists. The robustness of the iterative segmentation algorithm in terms of the initial contour provided to the algorithm was studied. The classification accuracy was found to depend on the initialization; however, the A_z value did not significantly deteriorate when the initial contour was scaled, rotated, or translated by a moderate amount. Future work includes verifying the results of this study by applying it to a larger and independent dataset, expanding the feature space by designing truly 3-D features, and combining the developed US characterization method with mammographic characterization methods. The observer performance study will also be performed to evaluate the effects of CAD on the characterization of breast masses by radiologists.

ACKNOWLEDGMENTS

This work was supported by USAMRMC Grant No. DAMD 17-01-1-0328 and by USPHS Grants No. CA095153 and No. CA91713. The content of this publication does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred. The authors are grateful to Charles E. Metz, Ph.D., for providing the LABROC program.

[a]Author to whom correspondence should be addressed. Berkman Sahiner, Ph.D., Department of Radiology, University of Michigan, 1500 E. Medical Center Drive, CGC B2102, Ann Arbor, Michigan 48109-0904. Telephone: (734) 647-7429; fax: (734) 615-5513; electronic mail: berki@umich.edu

[1] L. W. Bassett, T. H. Liu, A. I. Giuliano, and R. H. Gold, "The prevalence of carcinoma in palpable vs impalpable, mammographically detected lesions," AJR, Am. J. Roentgenol. **158**, 688–689 (1992).

[2] H. Opie, N. C. Estes, W. R. Jewell, C. H. Chang, J. A. Thomas, and M. A. Estes, "Breast biopsy for nonpalpable lesions: a worthwhile endeavor?," Am. Surg. **59**, 490–493 (1993).

[3] G. Hermann, C. Janus, I. S. Schwartz, B. Krivisky, S. Bier, and J. G. Rabinowitz, "Nonpalpable breast lesions: Accuracy of prebiopsy mammographic diagnosis," Radiology **165**, 323–326 (1987).

[4] F. M. Hall, J. M. Storella, D. Z. Silverstone, and G. Wyshak, "Nonpalpable breast lesions: recommendations for biopsy based on suspicion of carcinoma at mammography," Radiology **167**, 353–358 (1988).

[5] J. A. Baker, P. J. Kornguth, J. Y. Lo, and C. E. Floyd, "Artificial neural network: Improving the quality of breast biopsy recommendations," Radiology **198**, 131–135 (1996).

[6] Z. M. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," Acad. Radiol. **5**, 155–168 (1998).

[7] H.-P. Chan, B. Sahiner, M. A. Helvie, N. Petrick, M. A. Roubidoux, T. E. Wilson, D. D. Adler, C. Paramagul, J. S. Newman, and S. S. Gopal, "Improvement of radiologists' characterization of mammographic masses by computer-aided diagnosis: an ROC study," Radiology **212**, 817–827 (1999).

[8] A. T. Stavros, D. Thickman, C. L. Rapp, M. A. Dennis, S. H. Parker, and G. A. Sisney, "Solid breast nodules: Use of sonography to distinguish between malignant and benign lesions," Radiology **196**, 123–134 (1995).

[9] P. Skaane and K. Engedal, "Analysis of sonographic features in differentiation of fibroadenoma and invasive ductal carcinoma," AJR, Am. J. Roentgenol. **170**, 109–114 (1998).

[10] K. J. W. Taylor *et al.*, "Ultrasound as a complement to mammography and breast examination to characterize breast masses," Ultrasound Med. Biol. **28**, 19–26 (2002).

[11] B. Sahiner, G. L. LeCarpentier, H. P. Chan, M. A. Roubidoux, N. Petrick, M. M. Goodsitt, S. S. Gopal, and P. L. Carson, "Computerized characterization of breast masses using three-dimensional ultrasound images," Proc. SPIE **3338**, 301–312 (1998).

[12] D. R. Chen, R. F. Chang, and Y. L. Huang, "Computer-aided diagnosis applied to US of solid breast nodules by using neural networks," Radiology **213**, 407–412 (1999).

[13] K. Horsch, M. L. Giger, L. A. Venta, and C. J. Vyborny, "Computerized diagnosis of breast lesions on ultrasound," Med. Phys. **29**, 157–164 (2002).

[14] C. M. Chen, Y. H. Chou, K. C. Han, G. S. Hung, C. M. Tiu, H. J. Chiou, and S. Y. Chiou, "Breast lesions on sonograms: computer-aided diagnosis with nearly setting-independent features and artificial neural networks," Radiology **226**, 504–514 (2003).

[15] D. B. Downey, A. Fenster, and J. C. Williams, "Clinical utility of three-dimensional US," Radiographics **20**, 559–571 (2000).

[16] G. L. LeCarpentier, P. B. Tridandapani, J. B. Fowlkes, M. A. Roubidoux, A. P. Moskalik, and P. L. Carson, "Utility of 3-D ultrasound in discriminating and detection of breast cancer," RSNA EJ http://ej.rsna.org/ej3/0103-99.fin/titlepage.html, 1999.

[17] P. T. Bhatti, G. L. LeCarpentier, M. A. Roubidoux, J. B. Fowlkes, M. A. Helvie, and P. L. Carson, "Discrimination of sonographically detected breast masses using frequency shift color Doppler imaging in combination with age and gray scale criteria," J. Ultrasound Med. **20**, 343–350 (2001).

[18] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: active contour models," Int. J. Comput. Vis. **1**, 321–331 (1987).

[19] D. J. Williams and M. Shah, "A fast algorithm for active contours and curvature estimation," CVGIP: Image Understand. **55**, 14–26 (1992).

[20] L. D. Cohen, "On active contour models and balloons," CVGIP: Image Understand. **53**, 211–218 (1991).

[21] B. Sahiner, N. Petrick, H. P. Chan, L. M. Hadjiiski, C. Paramagul, M. A. Helvie, and M. N. Gurcan, "Computer-aided characterization of mammographic masses: Accuracy of mass segmentation and its effects on characterization," IEEE Trans. Med. Imaging **20**, 1275–1284 (2001).

[22] B. S. Garra, B. H. Krasner, S. C. Horri, S. Ascher, S. K. Mun, and R. K. Zeman, "Improving the distinction between benign and malignant breast lesions: The value of sonographic texture analysis," Ultrason. Imaging **15**, 267–285 (1993).

[23] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," IEEE Trans. Syst. Man Cybern. **SMC-3**, 610–621 (1973).

[24] Y. Zheng, J. F. Greenleaf, and J. J. Gisvold, "Reduction of breast biopsies with a modified self-organizing map," IEEE Trans. Neural Netw. **8**, 1386–1396 (1997).

[25] P. A. Lachenbruch, *Discriminant Analysis* (Hafner, New York, 1975).

[26] N. R. Draper, *Applied Regression Analysis* (Wiley, New York, 1998).

[27] C. E. Metz, B. A. Herman, and J. H. Shen, "Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," Stat. Med. **17**, 1033–1053 (1998).

[28] Y. Jiang, C. E. Metz, and R. M. Nishikawa, "A receiver operating characteristic partial area index for highly sensitive diagnostic tests," Radiology **201**, 745–750 (1996).

[29] B. Mesurolle, M. Ariche-Cohen, F. Mignon, J. Guinebretiere, A. Tardivon, and P. Goumot, "Small focal areas of acoustic shadowing in the breast," J. Clin. Ultrasound **30**, 88–97 (2002).

[30] K. Horsch, M. L. Giger, L. A. Venta, and C. J. Vyborny, "Automatic segmentation of breast lesions on ultrasound," Med. Phys. **28**, 1652–1659 (2001).

[31] G. Xiao, M. Brady, J. Noble, and Z. Yongyue, "Segmentation of ultrasound B-mode images with intensity inhomogeneity correction," IEEE Trans. Med. Imaging **21**, 48–57 (2002).

[32] A. Madabhushi and D. N. Metaxas, "Combining low, high-level and empirical domain knowledge for automated segmentation of ultrasonic breast lesions," IEEE Trans. Med. Imaging **22**, 155–169 (2003).

[33] I. Cohen, L. D. Cohen, and N. Ayache, "Using deformable surfaces to segment 3-D images and infer differential structures," CVGIP: Image Understand. **56**, 242–263 (1992).

Multi-modality CAD: Combination of computerized classification techniques based on mammograms and 3D ultrasound volumes for improved accuracy in breast mass characterization

Berkman Sahiner*, Heang-Ping Chan, Lubomir M. Hadjiiski, Marilyn A. Roubidoux,
Chintana Paramagul, Mark A. Helvie, Chuan Zhou
Department of Radiology, University of Michigan, Ann Arbor

ABSTRACT

Mammography and ultrasound (US) are two low-cost modalities that are commonly used by radiologists for evaluating breast masses and making biopsy recommendations. The goal of this study was to investigate computerized methods for combining information from these two modalities for mass characterization. Our data set consisted of 3D US images and mammograms of biopsy-proven solid breast masses from 60 patients. Thirty of the masses were malignant and 30 were benign. The US volume was obtained by scanning with an experimental 3D US image acquisition system. After computerized feature extraction from the 3D US images and mammograms, we investigated three methods (A, B and C) for combining the image features or classifier scores from different mammographic views and the US volumes. The classifier scores were analyzed using the receiver operating characteristic (ROC) methodology. The area A_z under the ROC curve of the classifier based on US alone was 0.88±0.04 for testing Two classifiers were designed using the mammograms alone, with test A_z values of 0.85±0.05 and 0.87±0.05, respectively. The test accuracy of combination methods A, B, and C were 0.89±0.04, 0.92±0.03, and 0.93±0.03, respectively. Our results indicate that combining the image features or classifier scores from the US and mammographic classification methods can improve the accuracy of computerized mass characterization.

Keywords: Computer-aided diagnosis, mammography, 3-D ultrasound, breast masses, lesion classification

1. INTRODUCTION

Masses are important indicators of malignancy in breast imaging. However, only a small percentage of breast masses evaluated on mammography and ultrasound (US) imaging are malignant.[1-4] Many benign masses may look suspicious enough on mammography and US examination for the radiologist to recommend biopsy. As a result, a large percentage of mass biopsies are performed for benign conditions. Benign biopsies not only causes patient discomfort and adds to medical costs, but also may result in scarring that can complicate the interpretation of future radiological exams. It is therefore very important to reduce the number of benign biopsies without missing any malignant masses.

Computer-aided diagnosis (CAD) can provide a consistent and reproducible second opinion to the radiologists, and has a potential to assist them in reducing benign biopsies. In recent years, considerable research effort has been devoted to the development of computerized feature extraction and classification methods for characterization of breast masses both on mammograms and US images.[5-11] Recent studies on the effect of computerized classification of breast masses on radiologists' characterization performance on mammograms indicate that radiologists' characterization may be significantly improved if they are aided by a well-trained CAD system.[7,12] Our recent studies also indicate that a similar improvement may be achieved when the radiologists rate the likelihood of malignancy of breast masses on 3D US images with computer aid. To our knowledge, no studies to date have investigated computerized classification of masses using combined information from 3D US images and mammograms. The purpose of this study was to evaluate the accuracy of different techniques for the combination of information from these two modalities in a computerized multi-modality breast mass classifier.

* berki@umich.edu, phone 734-647-7429, CGC B2102, 1500 E. Medical Center Dr., Ann Arbor, MI 48109-0904

Medical Imaging 2004: Image Processing, edited by
J. Michael Fitzpatrick, Milan Sonka, Proceedings of SPIE Vol. 5370
(SPIE, Bellingham, WA, 2004) · 1605-7422/04/$15 · doi: 10.1117/12.535868

2. METHODS

2.1 Feature extraction for classification of breast masses on mammograms

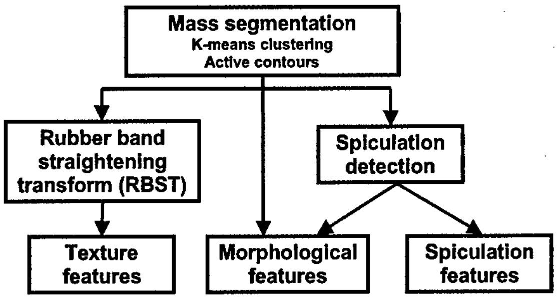


Figure 1: The block diagram of the mammographic mass segmentation and feature extraction method.

The block diagram of our segmentation and feature extraction method for mammographic masses is shown in Fig. 1. The mass segmentation method[13] consisted of two parts: K-means clustering and active contour segmentation. The purpose of the K-means clustering algorithm[14] was to classify each pixel in the region of interest (ROI) as an object (mass) or non-object (background) pixel, and to obtain an initial mass contour that would be refined using the active contour model. In our clustering algorithm, a pixel (i,j) was represented by a feature vector, whose components were obtained by filtering the original image by linear or nonlinear filters. We used three filtered images along with the original image to form the feature vectors. After the K-means algorithm clustered the pixels in the ROI into object and non-object classes, the boundary enclosing the object pixels was extracted to initialize our active contour segmentation method. In the active contour model, this initial boundary was iteratively deformed under internal and external forces in order to refine the mass boundary. The internal energy components were the continuity and curvature of the contour, as well as the homogeneity of the segmented object. The external energy components were the negative of the smoothed image gradient magnitude, and a balloon force that exerted pressure at a normal direction to the contour.

The active contour model was not suitable for the segmentation of spiculations because the curvature term in the model, that was essential for the regularity of the mass shape, prevents the contour from having sharp corners. For this reason, we designed an additional stage for the detection of spiculations. Spiculations on mammograms appear as linear structures with a positive image contrast, and they usually lie in a radial direction to the mass. As a result of their linearity, the gradient directions at image pixels on or close to the spiculation are more or less in the same orientation. In order to investigate whether a pixel (i_c,j_c) on the mass contour lies on the path of a spiculation, one can make use of this property as follows: In a search region S of the image, compute the statistics of the angular difference θ between the image gradient direction at image pixel (i,j), and the direction of the vector joining pixels (i_c,j_c), and (i,j). If a spiculation extends from the pixel (i_c,j_c), then θ will be close to $\pi/2$ whenever the image pixel (i,j) is on the spiculation. Therefore, the distribution of θ (as the image pixel (i,j) sweeps the search region S) will have a peak around $\pi/2$. We defined a spiculation measure based on the distribution of θ.[13] Figure 2.a plots the spiculation measure as (i_c,j_c) moves sequentially along the mass contour for the spiculated mass shown in Fig. 2.b, which was segmented using our active contour method. The locations of some of the local maxima in Fig. 2.a are also shown in Fig.2.b. It is observed that the maxima in Fig. 2.a correspond to locations where a linear structure extends from the mass. For the segmentation task, we computed the spiculation measure for a sequence of 30 contours. The first contour in the sequence was that provided by the active contour model. The following contours in the sequence were obtained by expanding the previous contour by one pixel, so that the spiculation measure was computed in a 30-pixel wide band around the mass. The resulting image in the 30-pixel-wide band around was named the spiculation likelihood map. Finally, the spiculation likelihood map image was used for segmenting the spiculations[13]. The resulting spiculation likelihood map and the final segmented masses are shown in Fig. 3.

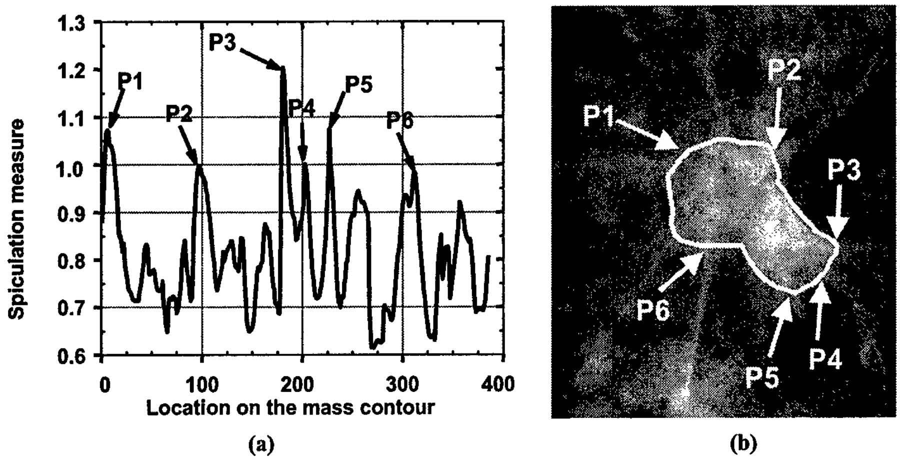


Figure 2: (a) The spiculation measure as a function of the location of the pixel (i,j) around the mass contour (b) The locations of some of the peaks in the spiculation measure profile shown in (a)

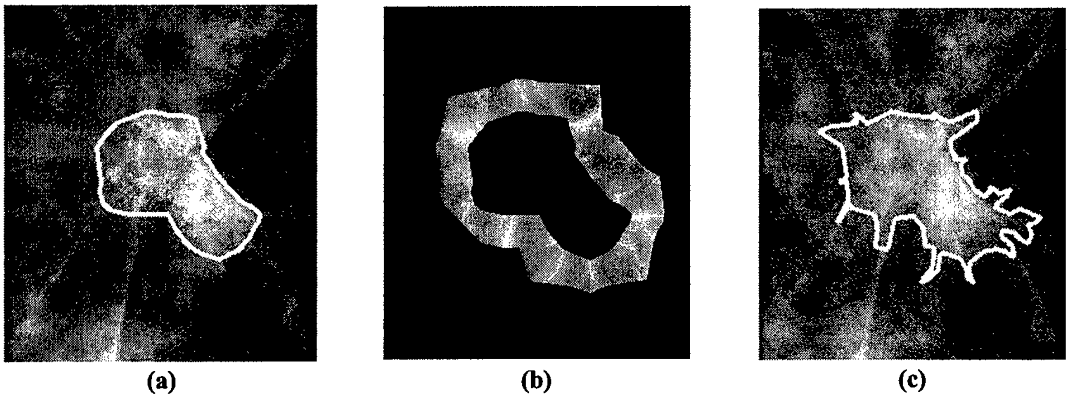


Figure 3: (a) The active-contour segmentation, (b) spiculation likelihood map, and (c) and the final segmentation result

After segmentation, morphological features were extracted from the segmented mass shapes. The extracted features included a Fourier descriptor, convexity, rectangularity, perimeter, contrast, circularity, perimeter-to-area ratio, area, normalized radial length (NRL) mean, NRL entropy, NRL area ratio, NRL standard deviation, and NRL zero crossing count.

Three spiculation features were extracted from the spiculation measure defined for the pixels along the boundary of the mass (an example of the spiculation measure is shown in Fig. 2.a). The first feature (AVG) was the average of the spiculation measure for all pixels on the mass boundary. The second feature (PERC_ABV) was the percentage of border pixels with a spiculation measure larger than $\pi/4$, and the third feature (AVE_ABV) was the average of the spiculation measure for those pixels with a spiculation measure larger than $\pi/4$. In addition, two spiculation features were extracted from the spiculation likelihood map defined in a band of pixels around the mass (an example of the spiculation likelihood map is shown in Fig. 3.b). The first feature (S_RATIO) was the percentage of the pixels within the ring-shaped area around the mass that were estimated to be spiculation pixels. The second feature (NS) was the product of the number of individual spiculations detected in the ring-shaped area and the S_RATIO feature.

We also extracted texture features from the band of pixels surrounding the mass. First, the band was transformed into Cartesian coordinates using the rubber-band straightening transform.[5] Then, RLS matrices, which describe the run-length statistics for each gray-level value in the image, were obtained from the vertical and horizontal gradient magnitudes of the RBST images. From each RLS matrix, five texture measures, namely, short runs emphasis, long runs emphasis, gray-level nonuniformity, run-length nonuniformity and run percentage, were extracted in the horizontal and vertical directions.[5]

2.2 Feature extraction for classification of breast masses on 3D US volumes

The block diagram of our segmentation and feature extraction method for mammographic masses is shown in Fig. 4.

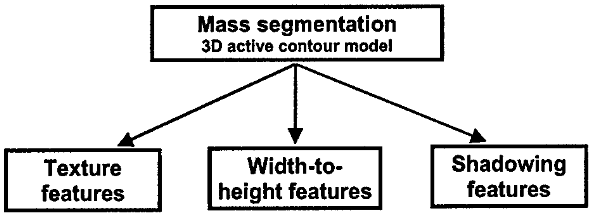


Figure 4: The block diagram of the segmentation and feature extraction method for masses on US volumes.

The 3D US images were acquired using an experimental system previously developed and tested at our institution.[15] The images were digitally stored and transferred to the workstation for processing. The biopsied mass in each volume was identified by an experienced radiologist using clinical US and mammographic images to confirm that the 3D images contained the suspicious mass. The radiologist was also asked to fit a 3D ellipsoid to the mass. The 3D ellipsoid was used to initialize the 3D active contour segmentation. In addition to the energy terms described above for our 2D active contour model, the 3D active contour model contained an additional curvature energy term related to the smoothness of the mass in the z-direction. Our results indicated that the use of this term increased the accuracy of 3D segmentation.[16] Figure 5 shows the result of the 3D segmentation algorithm for five consecutive slices of a mass.

After mass segmentation, morphological and texture features were extracted from each slice containing the mass. The morphological features included the width-to-height and posterior shadowing features. Both of these characteristics are known to be good indicators of malignancy.[17] The definitions of these features can be found in the literature.[8] The texture features were extracted from SGLD matrices derived from 2D slices of the 3D data set. It is known that the margin characteristics of a mass are very important for its characterization, and previous studies have indicated that texture features extracted from the mass margins are effective for classification.[18] For this reason, the texture features were extracted from two disk-shaped regions containing the boundary of each mass. Six texture feature measures that are invariant under linear, invertible gray scale transformations were extracted. More details about texture features used in this study can be found in the literature.[8]

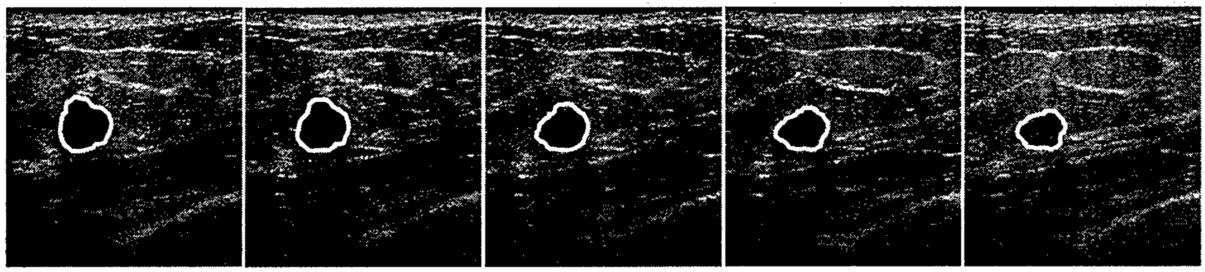


Figure 5: The segmentation result on five consecutive slices of a 3-D mass from our data set. This mass was seen on a total of 14 slices in the 3-D data set.

2.3 Combination methods

There are a number of strategies for designing a classifier that combines image information from different modalities. The first strategy is to design classifiers for each modality separately, and then to combine the classifier scores. The second strategy is to combine the features from the two modalities using a single classifier.

Considering the first strategy, there are again a number of classifier design options for each modality. For the 3D US volumes, the features were extracted from each slice, as described in the previous section. One can therefore consider the following two US classifiers for characterizing each mass case: (1) first develop a slice-based classifier that provides a malignancy score for each slice of each mass, and then combine the scores from different slices into a case-based classifier; and (2) first combine the feature vectors from different slices of the same mass into a case-based feature vector, and then design a case-based classifier. These two classifiers are called US_1 and US_2, respectively, in the following. For the mammograms, the features were extracted from each view containing the mass. For mammography, two classifiers parallel to US_1 and US_2 are: (1) first develop a view-based classifier, and then to combine the classifier scores from each view into a case-based classifier; and (2) first combine the feature vectors from different views into a case-based feature vector, and then design a case-based classifier. These two classifiers are called MAM_1 and MAM_2 in the following. Since the combination methods and the classifiers may not be linear, the classifiers designed using options (1) and (2) are, in general, different.

For each strategy or option, there are numerous methods to design a classifier and to combine features or classifier scores. In this study, we limited the classifier design method to linear discriminant analysis (LDA) with stepwise feature selection, and limited the combination operation to averaging. Among many alternatives for the overall multi-modality classifier design, we compared three: Methods A, B, and C. Methods A and B were based on designing classifiers for each modality separately, and then combining the classifier scores. In method A, we combined classifiers MAM_1 and US_2 defined above. In method B, we combined classifiers MAM_2 and US_2. Method C was based on first pooling the case-based feature vectors from mammography and US to define a larger feature space, and then designing a single classifier. These methods are summarized in Table 1.

Method A	Average the scores of the classifiers MAM_1 and US_2 described in the text
Method B	Average the scores of the classifiers MAM_2 and US_2 described in the text
Method C	Define case-based feature vectors for mammography and US by averaging the feature vectors from different views or slices. Pool these vectors to define a combined feature space, and design a single classifier in the combined feature space.

Table 1: A summary of Methods A, B, and C used for designing the multi-modality classifier

2.4 Data set

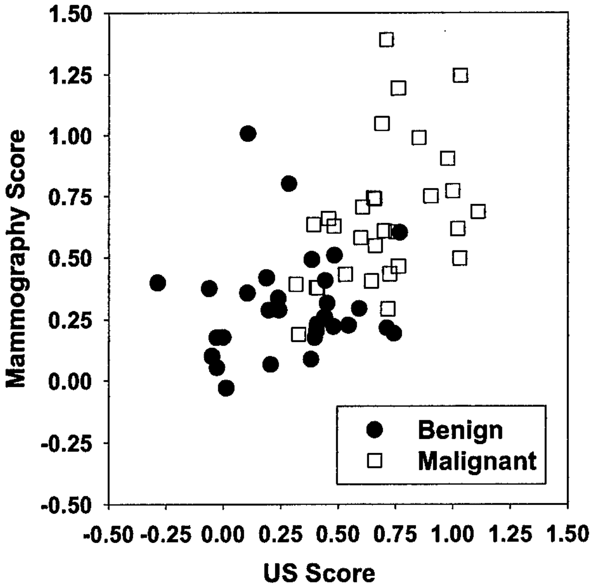
Our data set consisted of US volumes and mammograms from 60 patients who had a mammographically visible solid mass deemed suspicious or highly suggestive of malignancy. All patients underwent biopsy or fine needle aspiration. Thirty of the masses were malignant and 30 were benign. The biopsied mass on the mammograms and the US volumes was identified by an MQSA (Mammography Quality Standards Act) qualified radiologist using clinical images and case reports to confirm that the identified region contained the biopsied mass. The majority of the malignancies were invasive ductal carcinoma (N=26) and the majority of benign masses were fibroadenoma (N=14) and fibrocystic disease (N=4).

3. RESULTS

The classifiers were trained and tested using a leave-one-case-out method. The test classification accuracy of the single-modality classifiers MAM_1, MAM_2 and US_2 in terms of the area A_z under the receiver operating characteristic (ROC) curve is shown in Table 2. It can be observed that the accuracy of the classifiers MAM_1 and MAM_2 were lower than that of the US classifier; however, the difference did not reach statistical significance. The classification accuracy of the multi-modality classifiers A, B, and C are also shown in Table 2. It can be observed that the accuracies of all three multi-modality classifiers are higher than those of the single-modality classifiers. The two-tailed p-values obtained by comparing the A_z values of the single-modality and multi-modality classifiers are also listed in Table 2. The difference between the compared A_z values did not reach statistical significance, although there seems to be a strong trend that the multi-modality classifiers perform better.

Classifier	Case-based A_z	Two-tailed p value compared to method		
		A	B	C
Mammography alone (MAM_1)	0.85±0.05	0.08	0.07	0.09
Mammography alone (MAM_2)	0.87±0.05	>0.1	0.10	>0.1
US alone (US_2)	0.88±0.04	>0.1	>0.1	0.09
Combination method A	0.89±0.04			
Combination method B	0.92±0.03			
Combination method C	0.93±0.03			

Table 2: The classification accuracies of the single-modality classifiers (MAM1, MAM2, and US2), and multi-modality classifiers (Methods A, B, and C) that were investigated in this study.



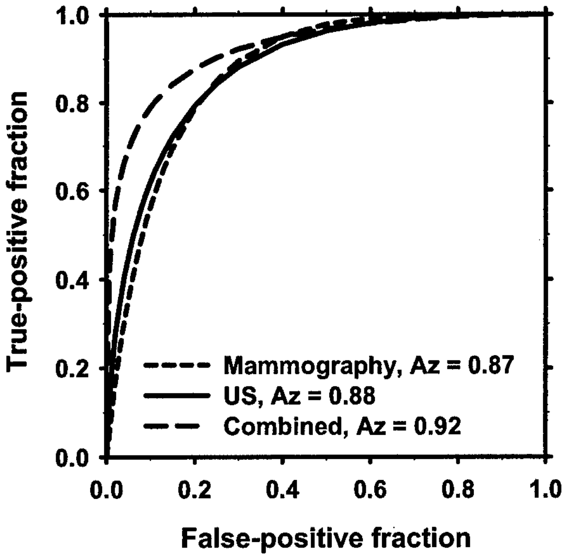


Figure 6: The distribution of the output scores of the classifiers MAM_2 and US_1 for the malignant and benign cases in our data set.

Figure 7: The ROC curves for the classifiers MAM_2, US_1 and their combination (method B).

In order to visually compare the distribution of the scores of individual masses under different modalities, the scatter plot of the MAM_2 and US_2 classifiers used in Method B are shown in Fig. 6 as an example. The scores from these two single-modality classifiers were not highly correlated (Pearson correlation coefficient = 0.54) so that the combination had a potential to achieve higher accuracy. The corresponding ROC curves, as well as the ROC curve for the combination, are shown in Fig. 7.

4. DISCUSSION AND CONCLUSION

To make a biopsy recommendation for a breast mass, radiologists routinely examine a patient's mammograms and the US images. These two modalities provide complementary information to the radiologists for more accurate diagnosis compared to that from a single modality. Despite the increased accuracy with these two modalities, the positive predictive value of biopsy recommendations is still low, and the radiologists may benefit from a multi-modality CAD system. In this study, we compared different methods for combining the information from mammograms and 3D US examinations for improving the accuracy of our computerized mass characterization system.

There are many ways in which computer-extracted diagnostic information from two modalities can be combined. In this study, we compared three methods. The Methods A and B were based on designing separate classifiers for US and mammography, and averaging the classifier scores from the two modalities at the end. Method C was based on pooling the extracted features from different modalities into a larger feature space, and designing a single classifier based on this feature space. Our results indicated that all three methods could improve the classification accuracy compared to those of single-modality classifiers. The difference, however, did not reach statistical significance, possible due to the small sample size.

We plan to enlarge our data set to investigate if the observed improvement with multi-modality CAD is generalizable, and to test the statistical significance of the difference between the multi- and single-modality classifiers. We will also investigate other multi-modality combination methods that were not tested in this study. Finally, we will perform observer performance studies to investigate the effect of our multi-modality computer classifier on radiologists' accuracy in characterizing malignant and benign masses.

ACKNOWLEDGMENTS

This work was supported in part by U.S. Army Medical Research Materiel Command grant DAMD17-01-1-0328 and by USPHS grants CA91713 and CA95153. The content of this paper does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned should be inferred.

REFERENCES

1. L. W. Bassett, T. H. Liu, A. I. Giuliano, and R. H. Gold, "The prevalence of carcinoma in palpable vs impalpable, mammographically detected lesions," *AJR* **158**, 688-689, 1992.

2. H. Opie, N. C. Estes, W. R. Jewell, C. H. Chang, J. A. Thomas, and M. A. Estes, "Breast biopsy for nonpalpable lesions: a worthwhile endeavor?," *American Surgeon* **59**, 490-493, 1993.

3. G. Hermann, C. Janus, I. S. Schwartz, B. Krivisky, S. Bier, and J. G. Rabinowitz, "Nonpalpable breast lesions: Accuracy of prebiopsy mammographic diagnosis," *Radiology* **165**, 323-326, 1987.

4. F. M. Hall, J. M. Storella, D. Z. Silverstone, and G. Wyshak, "Nonpalpable breast lesions: recommendations for biopsy based on suspicion of carcinoma at mammography," *Radiology* **167**, 353-358, 1988.

5. B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," *Medical Physics* **25**, 516-526, 1998.

6. R. M. Rangayyan, N. El-Faramawy, J. E. L. Desautels, and O. A. Alim, "Discrimination between benign and malignant breast tumors using a region-based measure of edge profile acutance," *In: Digital Mammography '96*, 213-218, Eds. K. Doi, M. L. Giger, R. M. Nishikawa and R. A. Schmidt, Elsevier, Amsterdam, 1996.

7. Z. M. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," *Academic Radiology* **5**, 155-168, 1998.

8. B. Sahiner, H. P. Chan, M. A. Roubidoux, M. A. Helvie, L. M. Hadjiiski, A. Ramachandran, G. L. LeCarpentier, A. Nees, C. Paramagul, and C. Blane, "Computerized characterization of breast masses on 3-D ultrasound volumes," *Medical Physics* (Accepted), 2003.

9. C. M. Chen, Y. H. Chou, K. C. Han, G. S. Hung, C. M. Tiu, H. J. Chiou, and S. Y. Chiou, "Breast lesions on sonograms: computer-aided diagnosis with nearly setting-independent features and artificial neural networks," *Radiology* **226**, 504-514, 2003.

10. D. R. Chen, R. F. Chang, and Y. L. Huang, "Computer-aided diagnosis applied to US of solid breast nodules by using neural networks," *Radiology* **213**, 407-412, 1999.

11. K. Horsch, M. L. Giger, L. A. Venta, and C. J. Vyborny, "Computerized diagnosis of breast lesions on ultrasound," *Medical Physics* **29**, 157-164, 2002.

12. H.-P. Chan, B. Sahiner, M. A. Helvie, N. Petrick, M. A. Roubidoux, T. E. Wilson, D. D. Adler, C. Paramagul, J. S. Newman, and S. S. Gopal, "Improvement of radiologists' characterization of mammographic masses by computer-aided diagnosis: an ROC study," *Radiology* **212**, 817-827, 1999.

13. B. Sahiner, H.-P. Chan, N. Petrick, M. A. Helvie, and L. M. Hadjiiski, "Improvement of mammographic mass characterization using spiculation measures and morphological features," *Medical Physics* **28**, 1455-1465, 2001.

14. B. Sahiner, H. P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue on mammograms," *Medical Physics* **23**, 1671-1684, 1996.

15. P. T. Bhatti, G. L. LeCarpentier, M. A. Roubidoux, J. B. Fowlkes, M. A. Helvie, and P. L. Carson, "Discrimination of sonographically detected breast masses using frequency shift color Doppler imaging in combination with age and gray scale criteria," *Journal of Ultrasound in Medicine* **20**, 343-350, 2001.

16. B. Sahiner, A. Ramachandran, H. P. Chan, L. M. Hadjiiski, M. A. Roubidoux, M. A. Helvie, C. Paramagul, A. Nees, C. Blane, N. Petrick, et al., "Three-dimensional active contour model for characterization of solid breast masses on three-dimensional ultrasound images," *Proceedings of the SPIE - Medical Imaging* **5032**, 405-413, 2003.

17. A. T. Stavros, D. Thickman, C. L. Rapp, M. A. Dennis, S. H. Parker, and G. A. Sisney, "Solid breast nodules: Use of sonography to distinguish between malignant and benign lesions," *Radiology* **196**, 123-134, 1995.

18. Y. Zheng, J. F. Greenleaf, and J. J. Gisvold, "Reduction of breast biopsies with a modified self-organizing map," *IEEE Transactions on Neural Networks* **8**, 1386-1396, 1997.

Appendix 3

Computer-Aided Diagnosis of Malignant and Benign Breast Masses in 3D

Ultrasound Volumes: Effect on Radiologists' Characterization Accuracy

Authors:

Berkman Sahiner, Ph.D.
Heang-Ping Chan, Ph.D.
Marilyn A. Roubidoux, M.D.,
Lubomir M. Hadjiiski, Ph.D.
Mark A. Helvie, M.D.
Chintana Paramagul, M.D.
Janet Bailey, M.D.
Alexis Nees, M.D.
Caroline Blane, M.D.

Institutional Affiliations:

Department of Radiology
The University of Michigan
CGC B2102,
1500 East Medical Center Drive
Ann Arbor, MI 48109-0904

Corresponding Author:

Berkman Sahiner, Ph.D.
Phone: 734-647-7429
Fax: 734-615-5513
Email: berki@umich.edu

Grant Information:

This work was supported by USAMRMC grant DAMD17-01-1-0328 and by USPHS grants CA095153 and CA91713.

RSNA Presentation:

This study was partly presented at the RSNA 2003 meeting (Abstract #504)

Type of manuscript:

Original research

Computer-Aided Diagnosis of Malignant and Benign Breast Masses in 3D Ultrasound Volumes: Effect on Radiologists' Characterization Accuracy

Type of manuscript:
Original research

ABSTRACT

Purpose: We have previously developed an automated computer classifier for characterization of malignant and benign breast masses in 3D ultrasound volumes. The purpose of this study was to investigate if computer aided diagnosis (CAD) using this classifier would improve radiologists' accuracy.

Materials and Methods: Our data set contained 3D ultrasound volumes from 102 cases of biopsy-proven breast masses (46 benign and 56 malignant). The 3D ultrasound volumes were recorded digitally while the transducer was translated across the lesion. A computer algorithm was designed to automatically delineate the mass boundaries and extract features based on the segmented mass shapes and margins. The features were merged into a malignancy score using a computer classifier. Five experienced radiologists participated as observers. Each radiologist read the cases first without CAD, immediately followed by reading with CAD. The observers' rating data were analyzed using receiver operating characteristic (ROC) methodology. The statistical significance of the differences was estimated by the Dorfman-Berbaum-Metz method and with the Student's two-tailed paired t test.

Results: Without CAD, the five radiologists had an average area under the ROC curve, A_z, of 0.84 (range: 0.81 to 0.87). With CAD, their average A_z increased significantly ($p=0.006$) to 0.90 (range: 0.86 to 0.93). Using a 2% likelihood of malignancy as the threshold for biopsy recommendation, the average sensitivity of the radiologists increased from 96% to 98% with CAD, while their average specificity for this data set decreased from 22% to 19%. If a biopsy recommendation threshold could be chosen such that the sensitivity were maintained at 96%, the specificity would increase to 46% with CAD.

Conclusion: A well-trained computer algorithm may improve radiologists' accuracy in characterizing breast masses as malignant and benign on ultrasound images.

Key Words: Computer-Aided Diagnosis, ROC Observer Study, Classification, Ultrasound, Malignancy.

INTRODUCTION

In current clinical practice, the positive biopsy rate for breast cancer is about 15-30% (1-3). To reduce patient anxiety and morbidity, as well as to decrease health care costs, it is desirable to reduce the number of benign biopsies without missing malignancies. Previous studies on mammography have shown that radiologists' accuracy in distinguishing malignant from benign masses can significantly improve when they use a well-trained computer-aided diagnosis (CAD) system as a second opinion. Chan et al. (4) performed an observer performance study in which six Mammography Quality Standards Act (MQSA) qualified radiologists rated the likelihood of malignancy of 238 biopsy-proven masses on a scale of 1 to 10 without and with a computer aid. The area A_z under the receiver operating characteristic (ROC) curve for the six radiologists ranged from 0.79 to 0.92 without CAD, and improved to 0.87-0.96 with CAD. The improvement was statistically significant ($p=0.022$). When the same radiologists read a subset of 76 paired views, their A_z was again significantly improved by CAD ($p=0.007$). In a study by Huo et al. (5) for differentiation of malignant and benign masses on 110 mammographic cases, the use of CAD improved the average A_z value of 12 radiologists from 0.93 to 0.96 ($p<0.001$). Recently, Hadjiiski et al. (6) found that CAD can improve significantly ($p=0.001$) radiologists' accuracy from an A_z in a range of 0.74-0.88 without CAD to a range of 0.76-0.92 reading sequentially with CAD when they interpret temporal pairs of masses on mammograms.

Ultrasound (US) is an important imaging modality for characterization of breast masses. For the differentiation of simple cysts from other lesions, interpretation of US images by experienced breast radiologists results in an accuracy close to 100% (7). In current clinical practice, if a palpable or mammographically suspicious mass cannot be confidently categorized as a cyst in US examination, it is often recommended for biopsy. Several recent studies (8-10) have indicated that the improvement in US imaging technology and the expert interpretation by radiologists may make it possible to characterize solid breast masses as malignant and benign with high accuracy. In a recent publication, Taylor et al. (10) reported that the addition of US evaluation to mammography alone increased the specificity in their data set of 761 biopsy-proven masses from 51.4% to 63.8%, while slightly increasing the sensitivity from 97.1% to 97.9%.

Several groups of researchers have been developing methods for computerized characterization of masses on 2-dimensional US images (11-14). We have recently developed an automated computer classifier for differentiation of malignant and benign breast masses in 3-dimensional (3D) US volumes (15). The purpose of this study was to investigate the effect of our computer classifier on radiologists' accuracy in discriminating between malignant and benign masses using 3D volumetric ultrasound images. Both the radiologists and the CAD algorithm analyzed 3D volumetric images of the masses which had been saved as cine-loops. To our knowledge, this is the first observer study to evaluate the impact of a CAD algorithm designed for 3D US images on radiologists' accuracy.

MATERIALS AND METHODS

Data Set

The data collection protocol was approved by our Institutional Review Board prior to the commencement of the study. Individual patient informed consent was obtained from all subjects. The group of imaged patients consisted of 130 consecutive patients who agreed to have a 3D breast US examination between 1998 and 2002. Eligibility criteria for subjects included women of any age who had a sonographic mass deemed suspicious or highly suggestive of malignancy, and who were scheduled for biopsy or fine needle aspiration. Twenty-eight patients from this study group were excluded as follows: patients who had prior biopsy in the same region of the breast, those with simple cysts, scans which were deemed technically unsuccessful because of motion or other artifacts, masses which were incompletely imaged in any dimension because of large size or eccentric position in the scan. Thus our study group consisted of 102 patients (average age: 51 years). Based on biopsy or fine needle aspiration results, 56 masses were malignant and 45 were benign. One of the masses resolved after imaging, and the patient was cancer-free after three year follow-up. Forty-three of the malignancies were invasive ductal carcinoma, 5 were invasive lobular carcinoma, 3 were ductal carcinoma in-situ, one was medullary carcinoma, and 4 were other invasive carcinoma. Of the biopsy-proven benign masses, 18 were fibroadenoma, 12 were fibrocystic disease, 8 were cyst, 2 were fat necrosis, 2 were scar tissue, one was fibrosis, one was granuloma, and one was other benign breast tissue. The mean lesion diameter was 1.28 cm (standard deviation = 0.78 cm).

The 3D US data were acquired using an experimental system that was previously developed and tested at our institution (16, 17). The 3D system consisted of a commercially available GE Logiq 700 (Milwaukee, WI) US scanner with an M12 linear array transducer, a mechanical transducer guiding system, and a computer workstation. The linear array transducer was operated at 11 MHz. The technologist was free to set the focal distance and the overall gain adjustment to obtain the best possible image. Before 3D image acquisition, the technologist used clinical US and mammogram images and reports to identify the suspicious mass. During 3D image acquisition, the technologist manually translated the transducer linearly in the cross-plane, or the z-direction, while the image acquisition system recorded 2D B-mode images in the image scan plane (x-y plane). The 2D images were obtained at approximately 0.5 mm incremental translations, which were measured and recorded using a translation sensor. The scanned breast region measured typically 4.5 cm long by 4.0 cm wide by 4.0 cm deep. The typical in-slice pixel size was approximately 0.11 mm X 0.11 mm.

The B-mode images were recorded into a buffer in the US scanner. After data acquisition, the images and the position data were transferred digitally to a workstation, where individual planes were cropped and stacked to form a 3D volume. The biopsy-proven mass in each volume was identified by an MQSA (Mammography Quality Standards Act) qualified radiologist, referred to as RAD0 in the following, using clinical US and mammographic images to confirm that the 3D images contained the mass of interest and showed the mass in its entirety.

Computerized Classification of Masses in US Volumes

The details of our CAD system developed for the classification of masses in 3D US volumes can be found in the literature (15). A summary of the method is provided below.

The first step of the CAD system involved the extraction of the mass boundaries in the 3D volume, i.e., mass segmentation. Automated segmentation of breast masses on

US images is a difficult task because of image speckles, posterior shadowing, and the variations of the gray level both within the mass and in the normal breast tissue. We developed a 3D active contour model for segmentation. The active contour model combined the prior knowledge about the relative smoothness of the 3D mass shape in US volume with the information in the image data. An example of the segmented mass slices for a malignant mass is shown in Figure 1.

After mass segmentation, image features were extracted from the mass and its margins for classification. Our feature space consisted of width-to-height ratio, posterior shadowing, and texture descriptors. The mass shape in terms of relative width to height was described by the ratio of the widest cross section of the automatically segmented lesion shape to the tallest cross section. Posterior shadowing features were defined in terms of the normalized average gray-level values in strips posterior to the mass. Texture features were extracted from two disk-shaped regions containing the boundary of each mass, as well as presumably mass and normal tissue adjacent to the boundary of the mass. These regions followed the contour determined by the active contour model. An illustration of the regions used for computing the posterior shadowing and texture features is shown in Figure 2. For additional details about the feature definitions, please refer to the Appendix.

The features described above were extracted from each slice of the US volume containing a mass to define slice-based features. For a given mass, features extracted from different slices were combined to define case-based features. Linear discriminant analysis (LDA) with stepwise feature selection (18) was applied to the case-based feature vectors to obtain computer-estimated malignancy scores. A leave-one-case-out resampling method (19) was used for training and testing of the classification system. The test scores obtained by the leave-one-out partitioning method were used as the malignancy scores in the observer performance study. Two Gaussian functions were fitted to the distributions of the malignancy scores of the benign and malignant classes separately, and were used in the observer performance study as described below.

Observer Performance Study

Five radiologists (RAD1-RAD5), different from the one who was involved in data set collection (RAD0), participated as observers. The radiologists RAD1-RAD5 had an average of 13 years of experience in mammographic and breast US interpretation (range: 3-26 years) in practice in an academic radiology department at a National Cancer Institute-designated comprehensive cancer center. They were all MQSA qualified. Four were fellowship-trained in breast imaging, and one had 26 years of experience in breast imaging. At our department, about 4300 breast US examinations are performed annually.

An interactive graphical user interface (GUI), shown in Figure 3, was developed to facilitate the navigation through 3D volumes, and to adjust the window and level of the displayed images. The location of the mass center, as determined by RAD0, was displayed on each slice, so that all the radiologists would rank the same mass if more than one mass could be seen in the volume.

During the experiment, an observer first read a case without CAD. This involved assessing mass characteristics in six categories such as shape, margins, echogenicity, and through transmission using the GUI, and providing an estimate of the likelihood of malignancy (LM) for the case on a scale of 0 to 100%. The complete list of the categories and the descriptors within each category are shown in Table 1. A button corresponding to an LM rating of 0% was provided for benign masses, and another button corresponding to LM ratings of less than 2% was provided for probably benign masses. This second button was set to correspond to the ACR-BIRADS category 3 (probably benign finding) for which short-interval follow-up is recommended (20). The radiologists used a slide bar to enter their ratings between 3% and 100%. The discrete buttons facilitate the selection of these LM ratings more precisely for the benign and probably benign masses because our previous experiences indicate that the uncertainty of selecting ratings on a slide bar by observers can be much greater than 2%. The observers were reminded at the beginning of the study that if they rated a mass as having larger than 2% of LM, it would indicate that they would recommend the mass for biopsy (20, 21).

We used a two-step sequential reading design. The radiologist first read the US volume without CAD, and rendered an estimate of the LM. The estimate without CAD was stored in a computer file, and the radiologist was not able to modify it after seeing the computer results. Immediately after reading without CAD, the computer-estimated malignancy score for the case was displayed on the screen, and the radiologist rendered an estimate of the LM with CAD. The computer's malignancy score is on a relative rating scale and cannot be easily converted to the likelihood of malignancy of the masses. We therefore linearly mapped and rounded the computer's malignancy score to an integer between 1 and 10 before displaying the score on the GUI. In order to provide a reference of the computer performance to the radiologists, the fitted Gaussian distributions to the computer scores for the malignant and benign classes were also displayed on the interface. The radiologists had the option to keep their original malignancy rating, or change it using the slide bar after taking into consideration the computer's opinion.

There was no time limit for the radiologists. The case reading order was randomized for each radiologist. In order to reduce the effect of fatigue on the radiologists' performance, the data set was read in three separate sessions by each radiologist. Before participating in the study, the radiologists were trained on five cases that were not part of the test data set. They were familiarized with the study design, the functions on the GUI, and the computer's relative malignancy rating scale during the training session.

Data Analysis

There is no ground truth for the mass characteristics such as echogenicity and through transmission, since they are judged subjectively by radiologists. To summarize the assessments of the mass characteristics, a "majority assessment" for each category was determined according to the majority rule by the six radiologists (RAD0-RAD5). The majority rule determined which one of the descriptors was selected by the largest number of radiologists. For example, if one radiologist described the echogenicity characteristics of a mass as hypoechoic, three as markedly hypoechoic, one as anechoic, and one as heterogeneous, the majority assessment for echogenicity of the mass would be markedly hypoechoic. When there was a tie between two descriptors, we used the descriptor chosen by RAD0, who was very familiar with the cases due to the role in data collection, as the tie-breaker. If there was a tie, and the original descriptor provided by RAD0 was not one of the descriptors that were tied, RAD0 was asked to re-read the images and choose one of the tied descriptors.

The LM ratings of the radiologists with and without CAD were analyzed using ROC methodology. The area under the ROC curve, A_z, and the partial area index above a sensitivity of 0.9, $A_z^{(0.9)}$ (22) were used as the accuracy measures. For an individual radiologist, the significance of the change in accuracy with CAD was also analyzed using ROC methodology. For the group of five radiologists, the significance of the change in accuracy with CAD was tested using the Dorfman-Berbaum-Metz (DBM) multi-reader multi-case (MRMC) methodology (23) and also using Student's two tailed paired t-test. The sensitivity and specificity of each radiologist with and without CAD were compared using an LM rating of 2% as the threshold above which biopsy would be recommended (20, 21).

In addition to analyzing the change with CAD in the number of cases for which the LM rating moved across the biopsy threshold of 2%, we also examined the number of cases for which the CAD resulted in a substantial change in the LM rating. We defined a substantial change as an absolute value difference of larger than or equal to 5 between LM ratings with and without CAD. The substantial decreases and increases in the ratings of malignant and benign cases were examined. For each mass, we also averaged the changes in the LM ratings for the five radiologists, and compared how CAD changes the average LM ratings for malignant and benign masses.

RESULTS

The majority assessments for the masses in our data set are shown in Table 1. A total of 96 masses were categorized as solid according to the majority rule. Five masses were categorized as complex cysts, and one as a simple cyst by three or more radiologists. One mass that was categorized as a complex cyst was malignant, and the remaining five non-solid masses were benign. The most common margin descriptor for malignant masses was ill-defined (46%), and that for benign masses was circumscribed (59%). Most of the malignant masses had irregular shape (59%) and most of the benign masses had oval shape (70%). Most of the masses (76% of benign masses and 64% of malignant masses) were categorized as hypoechoic. Calcifications were seen in 2% of benign masses and 25% of malignant masses.

The ROC curves of the five radiologists without and with CAD are shown in Figures 4a and 4b. The A_z values without CAD were in the range between 0.81 to 0.87 without CAD, and 0.86 to 0.93 with CAD. The average ROC curves for the radiologists with and without CAD were derived from the average a and b parameters, which were defined as the means of the individual radiologist's a and b parameters for the fitted ROC curves shown in Figure 5. The average ROC curves are shown in Figure 5 along with the test ROC curve of the computer classifier, which had an A_z value of 0.92.

Table 2 shows the individual radiologist's A_z and $A_z^{(0.9)}$ values with and without CAD, and the Student's two-tailed p-values for the change in both accuracy measures with CAD. Table 3 lists the average A_z and $A_z^{(0.9)}$ values, and the corresponding two-tailed p values estimated using the DBM method or the Student's paired t-test. The average A_z value improved from 0.84 to 0.90 with CAD. The improvement in A_z was statistically significant for the individual radiologists, as well as for the group of five radiologists. The average $A_z^{(0.9)}$ value improved from 0.30 to 0.47 with CAD. When the five radiologists were analyzed as a group, the improvement in $A_z^{(0.9)}$ was statistically significant; however, when analyzed individually, the improvement in $A_z^{(0.9)}$ did not reach statistical significance for three out of the five radiologists.

The sensitivity and specificity of each radiologist with and without CAD at an LM threshold of 2% are listed in Table 4. On the average, the radiologists' sensitivity increased from 96% to 98% with CAD, at the cost of a decrease in specificity from 22% to 19%. Three of the radiologists showed an increase in sensitivity while two maintained a sensitivity of 100%. The specificity of three radiologists decreased with CAD, while one radiologists' specificity increased and one did not show any change. Table 4 also shows the sensitivity and specificity for each radiologist if the LM threshold were to be adjusted to 7% when they read with CAD, for which the average sensitivity would remain at 96% (same as that without CAD) while the average specificity would increase to 46%.

With 102 cases and five radiologists, we had a total of 510 pairs of LM ratings with and without CAD. Figure 6 shows a histogram of the change in the radiologists' LM ratings with CAD for these 510 readings. The radiologists did not change their LM rating substantially (i.e., within 5) with CAD in 64% (326/510) of the readings. For malignant masses, the ratings were substantially increased for 34% (95/280) and decreased for 7% (19/280) of the readings. For benign masses, the ratings were substantially increased for 14% (32/230) and decreased for 17% (38/230) of the readings.

Figure 7 shows the histogram of the mean change in the LM ratings for malignant and benign masses. To obtain the mean change for a mass, the changes with CAD from five radiologists were averaged. To statistically evaluate the change for malignant and benign masses, we performed one-sample t-tests on the mean changes. For benign masses, the decrease in the average LM rating was 0.77, which did not achieve statistical significance (two-tailed $p=0.51$). The increase in the average LM rating of malignant masses was 5.59, which was statistically significant (two-tailed $p<0.0001$).

As described at the beginning of this section, 96 masses were categorized as solid by the majority rule. To investigate how the radiologists performed for this subset of cases, we applied ROC analysis to this subset by excluding cases that were categorized as complex or simple cysts. The average A_z values without and with CAD for this subset were 0.84 and 0.90, respectively, unchanged from the set of 102 cases. The improvements in A_z for the individual radiologists as well as for all radiologists as a group were statistically significant ($p<0.05$).

DISCUSSION

Our results indicate that the CAD algorithm used in this study was able to assist even expert breast imaging radiologists in characterizing masses on 3D US volumes. At our institution, all clinical breast US examinations are performed by breast imaging radiologists, not sonographers, and therefore they are particularly experienced in assessing whole volume images. Nevertheless, our CAD system could improve their accuracy in terms of the A_z and $A_z^{(0.9)}$ values. The average A_z value improved significantly ($p=0.005$) from 0.84 to 0.90, and the average $A_z^{(0.9)}$ value improved significantly (p=0.015) from 0.30 to 0.47. The area under the ROC curve for the computer classifier (A_z =0.92) was higher than those of all radiologists without CAD in the study. With CAD, all radiologists showed a significant improvement in their A_z values, and one radiologist's A_z value surpassed that of the computer classifier.

All of the masses in this study were deemed suspicious or highly suggestive of malignancy at the time of data collection. During the observer experiment, 96/102 (94%) of the masses were assessed as solid according to the majority rule. When the analysis was limited to this subset of solid masses, the A_z values with and without CAD, and the significance of the improvement with CAD were essentially unchanged compared to the results with the entire data set of 102 cases. This implies that CAD would be helpful for the interpretation of the more difficult category of solid masses.

The effect of CAD was mixed when measured in terms of the radiologists' sensitivity and specificity values at the threshold of biopsy recommendation (LM of 2%). With CAD, the average sensitivity of the five radiologists increased from 96% to 98%, while their average specificity for this data set decreased from 22% to 19%. Without the benefit of the malignancy ratings recorded in the observer experiment, it would not have been possible to ascertain whether these changes in the specificity and sensitivity reflect only a shift in decision threshold along the same ROC curve. Our malignancy rating data strongly suggests that this is not the case, as evidenced by the significant improvement in the ROC curves. Since all lesions except one in our data set underwent biopsy or fine needle aspiration after clinical imaging, the relatively low specificity of the radiologists with or without CAD is not unexpected.

The ultimate clinical utility of a CAD system that results in an increased sensitivity at the cost of decreasing specificity depends on a cost/benefit analysis of the different correct and incorrect decisions. Alternatively, by appropriate training, it may be possible to translate the benefits with CAD into biopsy decisions that surpass unaided reading in terms of both sensitivity and specificity, or an improvement in specificity without reducing sensitivity. For example, for our data set, if the threshold for biopsy with CAD could be changed to an LM rating of 7%, the average specificity with CAD would have been improved to 46%, compared to 22% without CAD, while the average sensitivity would remain at 96% as noted above.

Since the "cost" of failing to biopsy a malignant lesion is much greater than that of a benign biopsy, it can logically be expected that radiologists may tend to use the CAD system to confirm and increase their LM estimate of malignant lesions while not easily reducing the LM estimate of low suspicion lesions. This will result in an overall increase in radiologists' LM ratings. Figure 6 suggests that this is indeed the case in our study. While the ratings for malignant masses demonstrated a stronger trend to increase than to decrease with CAD, the ratings for benign masses did not show a strong trend either way. It is also noted that the radiologists' ratings showed little or no change (less than 5%) for a large percentage (64%) of the masses. It therefore appears that radiologists tend to be very conservative in downgrading the LM of a lesion. As a result, the observed improvement in the radiologists' accuracy in this study was obtained mainly from an increase in the LM ratings of malignant masses. This led to an increase in sensitivity and a slight decrease in specificity. However, since the ROC curves of all radiologists did improve with CAD, there is a potential that the radiologists can adjust their decision thresholds along the higher ROC curves and thus increase the sensitivity as well as the specificity. Alternatively, it may be possible to convince them to reduce the LM ratings of masses that the CAD system rates as very low suspicion, thus improving the specificity. These improvements may be realized after radiologists accumulate experiences and increase their confidence with the use of CAD.

The assessment of mass characteristics shown in Table 1 helped us better identify the properties of our data set. It was reported that a systematic analysis of the characteristics of breast lesions guided by a checklist could improve radiologists' diagnostic accuracy (24). The list of mass descriptors collected in this study is similar to that in the ultrasound BI-RADS lexicon recently published by the American College of Radiology (20). However, since the BI-RADS lexicon for breast US had not been published at the time of the study, the descriptors are not exactly the same. A study by Rahbar et al. (25) investigated the correlation of US features and tissue diagnosis. Similar to our study, the most common shape and margin descriptors for benign masses in that study were round or oval shape and circumscribed margins, and the most common shape and margin descriptors for malignant masses were irregular shape and ill-defined margins.

A number of research groups have been developing CAD systems for breast mass characterization on US images in recent years. (12-15). Chen et al. (13) used morphological features extracted from hand-segmented mass boundaries on 2D US images to design a nearly setting-independent classifier. Using an automated segmentation method, Horsch et al. (14) obtained an A_z value of 0.87 in the task of differentiating all malignant and benign lesions (N=400) in their 2D US data set, and

0.82 in the task of differentiating the subset of malignant and benign solid lesions (N=276). Sahiner et al. (15) designed a classifier based on features extracted from 3D US images, and found that the accuracy of the designed classifier in estimating the likelihood of malignancy of masses was similar to that of experienced radiologists when their performances were compared for the same set of images. These previous studies, therefore, indicate that computer classifiers can perform well for characterizing masses on US images, although it is not possible to directly compare the performances of the classifiers because they were tested on different data sets. However, we are aware of very few studies that investigated the effect of CAD for US mass characterization on radiologists' accuracy. Recently, Horsch et al. (26) found that the accuracy of both expert mammographers and community radiologists improved significantly when they read 2D US images with CAD. Our study differs from that by Horsch et al. in that 3D US images were used but our results reinforce the finding that experienced radiologists can benefit from reading US images with CAD.

The US images used for analysis by the CAD system in this study constituted a volume that contained the biopsy-proven mass, acquired using an experimental system. The radiologists in our observer study were asked to characterize the masses based on the same US volumes. In clinical practice, typically, these readers will interactively optimize the image quality by changing the probe angle, direction, and US scan settings for a given case. The images interpreted in our observer performance study were therefore different from those our radiologists routinely interpret. The potentially less than optimal image quality may have had a negative impact on their reading accuracy. To our knowledge, all CAD systems developed so far for breast US operate on static images, and therefore do not take advantage of the interactive nature of US imaging. The use of 3D volumes for CAD design may reduce this disadvantage by providing a more complete description of the mass compared to a few 2D images containing the mass. Similarly, interpretation of 3D US volumes by a radiologist may offer advantages compared to interpretation of only a few hardcopy images acquired by a US technologist, although interactive acquisition by a radiologist may still be the best approach. Although current CAD systems have been designed for off-line processing of recorded US images to facilitate algorithm development in the laboratory, it is conceivable that the processing may be sped up to real time or within seconds of the US exam by firmware implementation in the future to make it compatible with clinical operations.

Our study had a number of limitations. As described in the Introduction, one of the purposes of our CAD system was to help radiologists reduce the benign biopsy rate without affecting the sensitivity of breast cancer detection. Our data set therefore consisted of only masses that were recommended for biopsy or fine needle aspiration. However, if such a system were used prospectively, it may affect the management of cases that the radiologist would normally recommend for a follow-up. It is therefore important in the future to investigate the performance of the CAD system for masses that are not recommended for biopsy, and whose outcomes are known by follow-up. A second limitation is that all the cases in our data set were collected using the same US machine. Although we believe that our image processing methods will not depend strongly on small changes in image quality of the US images, the CAD system needs to be evaluated with images acquired using different US imaging systems to ensure its robustness against variations in image acquisition systems and parameters. A third limitation is that all the observers in our study were very experienced in breast imaging and US interpretation so that the effects of CAD on less experienced radiologists are still unknown. We believe that less experienced readers may benefit from CAD at least as much as the experienced radiologists, if not more. Fourth, our CAD system was trained and tested using a leave-one-case-out method. Although this is known as a nearly unbiased classifier design method (19), the performance of our CAD system needs to be evaluated using independent test sets in order to assure the generalizability of our approach. However, this study did reveal the potential benefits CAD may provide to the radiologists for the characterization of masses, given that a CAD system with the level of performance used in our study is available as a second opinion. Finally, radiologists generally combine information from US with that from mammograms to reach a diagnostic decision while the current study only used the information from US images. The effects of CAD on a combined US and mammogram evaluation remain to be investigated.

ACKNOWLEDGMENTS

This work was supported in part by U.S. Army Medical Research Materiel Command grant DAMD17-01-1-0328 and by USPHS grants CA095153 and CA91713. The authors are grateful to Charles E. Metz, Ph.D., for the LABMRMC program.

APPENDIX

Feature extraction

The feature vector for a given mass consisted of four width-to-height features, four posterior shadowing features, and 72 texture features.

The width-to-height features for a mass were the minimum, maximum, mean, and the standard deviation of the ratio of the width to the height of the segmented mass for each slice containing the mass. The width W and height H of the segmented mass in a slice were defined as the widest and the tallest cross-sections of the mass in that slice, respectively (Figure 2).

The posterior shadowing features for a mass were the minimum, maximum, mean, and the standard deviation of the feature extracted from each slice containing the mass. On a given slice, the posterior region of the mass was divided into n overlapping vertical strips as shown in Figure 2. The width of each strip was equal to $W/4$, and the height of the strip was equal to H. The strips were defined only posterior to the central $3W/4$ portion of the mass so that bilateral shadows that are sometimes associated with fibroadenomas could be avoided. Let P denote the mean grayscale value within the darkest posterior strip, and M denote the mean grayscale value within the segmented mass. The difference D between M and P defined how dark the US image is in the darkest posterior strip of the mass compared to the average within the mass. The posterior shadowing feature for the slice was defined as the normalized difference D/M.

The texture features were extracted from disc-shaped regions posterior and anterior to the mass. These equal-sized regions contained partly the interior portion of the mass and partly the mass margins. The total area of the anterior and posterior regions was equal to the area of the segmented mass. An example of the anterior disc-shaped region is shown in Figure 2. On each slice containing the mass, spatial gray level dependence (SGLD) matrices, $S(d, \theta)$ were extracted. The $(i,j)^{th}$ element of $S(d, \theta)$ is the relative frequency with which two pixels, one with gray level i and the other with gray level j, separated by a pixel pair distance d in a direction θ, occur in the image. In this study, three pixel pair distances, d=2, 4, and 6, and two pixel pair angles, θ=0° and 90° were used. On each slice, we therefore extracted six SGLD matrices from the anterior and six SGLD matrices from the posterior disc-shaped regions. From each SGLD matrix, six texture features were extracted. These features were information measures of correlation 1 and 2, entropy, difference entropy, sum entropy, and energy. The mathematical definitions of these features can be found in the literature (27). The texture feature vector extracted from a slice was therefore 72-dimensional. These vectors were averaged over all slices containing a mass to obtain the texture feature vector for the mass.

REFERENCES

1. Kopans DB. The positive predictive value of mammography. AJR 1992; 158:521-526.
2. Adler DD, Helvie MA. Mammographic biopsy recommendations. Curr. Op. Radiol. 1992; 4:123-129.
3. Brown ML, Houn F, Sickles EA, Kessler LG. Screening mammography in community practice: positive predictive value of abnormal findings and yield of follow-up diagnostic procedures. AJR 1995; 165:1373-1377.
4. Chan HP, Sahiner B, Helvie MA, Petrick N, Roubidoux MA, Wilson TE, Adler DD, Paramagul C, Newman JS, Gopal SS. Improvement of radiologists' characterization of mammographic masses by computer-aided diagnosis: an ROC study. Radiology 1999; 212:817-827.
5. Huo ZM, Giger ML, Vyborny CJ, Metz CE. Breast cancer: Effectiveness of computer-aided diagnosis - Observer study with independent database of mammograms. Radiology 2002; 224:560-568.
6. Hadjiiski LM, Chan HP, Sahiner B, Helvie MA, Roubidoux M, Blane C, Paramagul C, Petrick N, Bailey J, Klein K, Foster M, Patterson S, Adler D, Nees A, Shen J. Improvement of Radiologists' Characterization of Malignant and Benign Breast Masses in Serial Mammograms by Computer-Aided Diagnosis: An ROC Study. Radiology 2004:(in press).
7. Jackson VP. The role of US in breast imaging. Radiology 1990; 177:305-311.

8. Stavros AT, Thickman D, Rapp CL, Dennis MA, Parker SH, Sisney GA. Solid breast nodules: Use of sonography to distinguish between malignant and benign lesions. Radiology 1995; 196:123-134.

9. Skaane P, Engedal K. Analysis of sonographic features in differentiation of fibroadenoma and invasive ductal carcinoma. AJR 1998; 170:109-114.

10. Taylor KJW, Merritt C, Piccoli C, Schmidt R, Rouse G, Fornage B, Rubin E, Georgian-Smith D, Winsberg F, Goldberg B, Mendelson E. Ultrasound as a complement to mammography and breast examination to characterize breast masses. Ultrasound Med. Biol. 2002; 28:19-26.

11. Garra BS, Krasner BH, Horri SC, Ascher S, Mun SK, Zeman RK. Improving the distinction between benign and malignant breast lesions: The value of sonographic texture analysis. Ultrasonic Imaging 1993; 15:267-285.

12. Chen DR, Chang RF, Huang YL. Computer-aided diagnosis applied to US of solid breast nodules by using neural networks. Radiology 1999; 213:407-412.

13. Chen CM, Chou YH, Han KC, Hung GS, Tiu CM, Chiou HJ, Chiou SY. Breast lesions on sonograms: computer-aided diagnosis with nearly setting-independent features and artificial neural networks. Radiology 2003; 226:504-514.

14. Horsch K, Giger ML, Venta LA, Vyborny CJ. Computerized diagnosis of breast lesions on ultrasound. Med. Phys. 2002; 29:157-164.

15. Sahiner B, Chan HP, Roubidoux MA, Helvie MA, Hadjiiski LM, Ramachandran A, LeCarpentier GL, Nees A, Paramagul C, Blane CE. Computerized characterization of breast masses on 3-D ultrasound volumes. Med. Phys. 2004; 31:744-754.

16. Bhatti PT, LeCarpentier GL, Roubidoux MA, Fowlkes JB, Helvie MA, Carson PL. Discrimination of sonographically detected breast masses using frequency shift color Doppler imaging in combination with age and gray scale criteria. Journal of Ultrasound in Medicine 2001; 20:343-350.

17. Carson PL, Fowlkes JB, Roubidoux MA, Moskalik AP, A. G, Normolle D, LeCarpentier GL, Nattakom S, Helvie MA, Rubin JM. 3-D color Doppler image quantification of breast masses. Ultrasound Med. Biol. 1998; 24:945-952.

18. Draper NR. Applied regression analysis. New York: Wiley, 1998.

19. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. New York: Springer-Verlag, 2001.

20. American College of Radiology Breast Imaging Reporting and Data System Atlas (BI-RADS Atlas). Reston, VA: American College of Radiology, 2003.

21. Sickles EA. Nonpalpable, circumscribed, noncalcified solid breast masses: likelihood of malignancy based on lesion size and age of patient. Radiology 1994; 192:439-442.

22. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology 1996; 201:745-750.

23. Dorfman DD, Berbaum KS, Metz CE. ROC rating analysis: Generalization to the population of readers and cases with the jackknife method. Invest. Radiol. 1992; 27:723-731.

24. Getty DJ, Pickett RM, D'Orsi CJ, Swets JA. Enhanced interpretation of diagnostic images. Invest Radiol 1988; 23:240-252.

25. Rahbar G, Sie AC, Hansen GC, Prince JS, Melany ML, Reynolds HE, Jackson VP, Sayre JW, Bassett LW. Benign versus malignant solid breast masses: US differentiation. Radiology 1999; 213:889-894.

26. Horsch K, Giger ML, Vyborny CJ, Venta LA. Performance of computer-aided diagnosis in the interpretation of lesions on breast sonography. Acad. Radiol. 2004; 11:272-280.

27. Haralick RM, Shanmugam K, Dinstein I. Texture features for image classification. IEEE Trans. Sys. Man. and Cybern. 1973; SMC-3:610-621.

TABLES

Table 1.

Characteristics of the masses in our data set. Each characteristic of a mass was determined from the assessments by the six radiologists using a majority voting method, in which the descriptor that was selected by the largest number of radiologists was chosen.

Overall US impression	Shape	Margins	Echogenicity	Through transmission	Other features
Negative B: 0 (0) M: 0 (0)	Oval B: 32 (70) M:13 (23)	Circumscribed B: 27 (59) M: 3 (5)	Echogenic B: 0 (0) M: 0 (0)	Increased transmission B: 13 (28) M: 15 (27)	Taller than wide B: 1 (2) M: 11 (20)
Simple Cyst B: 1 (2) M: 0 (0)	Round B: 8 (17) M: 3 (5)	Spiculated B: 1 (2) M: 7 (13)	Isoechoic B: 5 (11) M: 3 (5)	Distal shadowing B: 11 (24) M: 20 (36)	Thin echogenic rim B: 2 (4) M: 1 (2)
Complex Cyst B: 4 (9) M: 1 (2)	Lobulated B: 2 (4) M: 7 (13)	Microlobulated B: 5 (11) M: 20 (36)	Hypoechoic B: 35 (76) M: 36 (64)	Neither B: 22 (48) M: 21 (38)	Ductal extension B: 0 (0) M: 3 (5)
Solid B: 41 (89) M: 55 (98)	Irregular B: 4 (9) M:33 (59)	Ill defined B: 13 (28) M: 26 (46)	Markedly hypoechoic B: 4 (9) M: 9 (16)		Calcifications B: 1 (2) M: 14 (25)
			Anechoic B: 1 (2) M: 1 (2)		Echogenic halo B: 1 (2) M: 2 (4)
			Heterogeneous B: 1 (2) M: 7 (13)		

Note — The numbers in parentheses are the percentages of the descriptors relative to the total number of benign and malignant masses in the data set. Benign (B): N=46, Malignant (M): N=56.

Table 2.

The area A_z under ROC curve, and the partial area index $A_z^{(0.9)}$ above a sensitivity of 0.9, for the characterization of the masses in the data set without and with CAD by the 5 radiologists.

Rad.	A_z			$A_z^{(0.9)}$		
No	No CAD	With CAD	p value*	No CAD	With CAD	p value*
1	0.83±0.04	0.89±0.03	0.0008	0.25±0.10	0.35±0.14	0.17
2	0.81±0.04	0.86±0.04	0.0005	0.14±0.08	0.23±0.12	0.13
3	0.87±0.03	0.91±0.03	0.0486	0.39±0.12	0.53±0.12	0.0747
4	0.82±0.04	0.93±0.02	0.0004	0.39±0.10	0.68±0.09	0.0008
5	0.83±0.04	0.90±0.03	0.0007	0.29±0.10	0.42±0.12	0.0323

Note —The A_z and $A_z^{(0.9)}$ values are the mean ± SD.

* The p value from the Student's two-tailed paired t-test for each radiologist is shown.

Table 3.

The average A_z and $A_z^{(0.9)}$ values without and with CAD for the five radiologists, obtained by using the average a and b parameters.

Accuracy measure	No CAD	With CAD	p value (DBM)	p value (paired t-test)
A_z	0.84	0.90	0.006	0.005
$A_z^{(0.9)}$	0.30	0.47	---	0.015

Note — The significance of the change in the A_z value with CAD for the group of five radiologists was estimated using both the DBM method and the Student's two-tailed paired t-test. The significance of the change in the $A_z^{(0.9)}$ value was estimated using the Student's two-tailed paired t-test.

Table 4.

The sensitivity and specificity for each radiologist.

Rad.	Sensitivity			Specificity		
No.	No CAD*	With CAD*	With CAD**	No CAD*	With CAD*	With CAD**
1	56 (100)	56 (100)	56 (100)	4 (9)	5 (11)	15 (33)
2	51 (91)	53 (95)	49 (88)	12 (26)	11 (24)	28 (61)
3	52 (93)	54 (96)	53 (95)	24 (52)	22 (48)	29 (63)
4	55 (98)	56 (100)	56 (100)	9 (20)	5 (11)	23 (50)
5	56 (100)	56 (100)	56 (100)	1 (2)	1 (2)	11 (24)
Avg.	54 (96)	55 (98)	54 (96)	10 (22)	9 (19)	21 (46)

Note — In each entry, the first number denotes the number of correctly classified lesions, and the number in parentheses denotes the percentage (i.e., sensitivity for the first three columns. and the specificity for the last three columns). The total numbers of malignant and benign lesions are 56 and 46, respectively.

* The columns entitled "No CAD*" and "With CAD*" show the sensitivity and specificity at the decision threshold of 2% likelihood of malignancy, without and with CAD, respectively.

** The columns entitled "With CAD**" show the hypothetical sensitivity and specificity with CAD at a decision threshold of 7% likelihood of malignancy, for which the average sensitivity would be the same as that without CAD (96%), but the average specificity would be increased to 46%.

CAPTIONS FOR ILLUSTRATIONS

Figure 1: Five slices containing a malignant mass and the result of computer segmentation.

Figure 2: For feature extraction, the width W and height H of the mass on a slice were defined as the widest and the tallest cross-sections of the mass in that slice, respectively. The mean gray level values within the overlapping posterior strips $R(i)$ and the segmented mass were used to define the posterior shadowing features. The disc-shaped regions for texture feature extraction followed the shape of the mass and contained partly the segmented mass and partly its margins. An example of the anterior disc-shaped region is shown as the gray area above the segmented mass.

Figure 3: The graphical user interface. The biopsy-proven lesion was marked by an arrow, which could be switched off when the radiologist assessed the mass. The interface allowed the users to navigate through the volume, and to adjust the contrast and brightness. The radiologists first provided their assessment for the mass in six categories, which were 1) overall US impression; 2) shape; 3) margins; 4) echogenicity; 5) through transmission; and 6) other features. They then provided a likelihood of malignancy rating without CAD. Finally, the computer's malignancy score for the mass was displayed and the radiologists had an option to revise their rating after taking into consideration the computer's opinion.

Figure 4: (a) The ROC curves of the five radiologists without CAD and (b) with CAD. The area under the ROC curve A_z and the partial area above a sensitivity threshold of 0.9 $A_z^{(0.9)}$ are shown in Table 2 for each radiologist.

Figure 5: The average ROC curves of the radiologists with and without CAD, and the ROC curve of the computer classifier. The average ROC curves were constructed by using the mean a and b values of the individual observers' ROC curves shown in Figure 4.

Figure 6: The histogram of the change in radiologists' ratings with CAD. For the majority of the masses (59% of malignant masses and 70% of benign masses) the change was in the range of -4 to 4. When the change in the scores with CAD was greater than or equal to the range of -5 to 5, the change was called substantial. For malignant masses, the ratings were substantially increased for an average of 34% (95/280) and decreased for 7% (19/280) of the readings. For benign masses, the ratings were substantially increased for 14% (32/230) and decreased for 17% (38/230) of the readings.

Figure 7: The histogram of the mean change in the LM ratings of radiologists with CAD. The mean change for a mass was computed by averaging the changes in the LM ratings for that mass over the five radiologists who participated in the study. For benign masses, the overall average LM rating decrease was 0.77, which did not achieve statistical significance (p=0.51). For malignant masses the overall average LM rating increase was 5.59, which was statistically significant (p<0.0001).

ILLUSTRATIONS

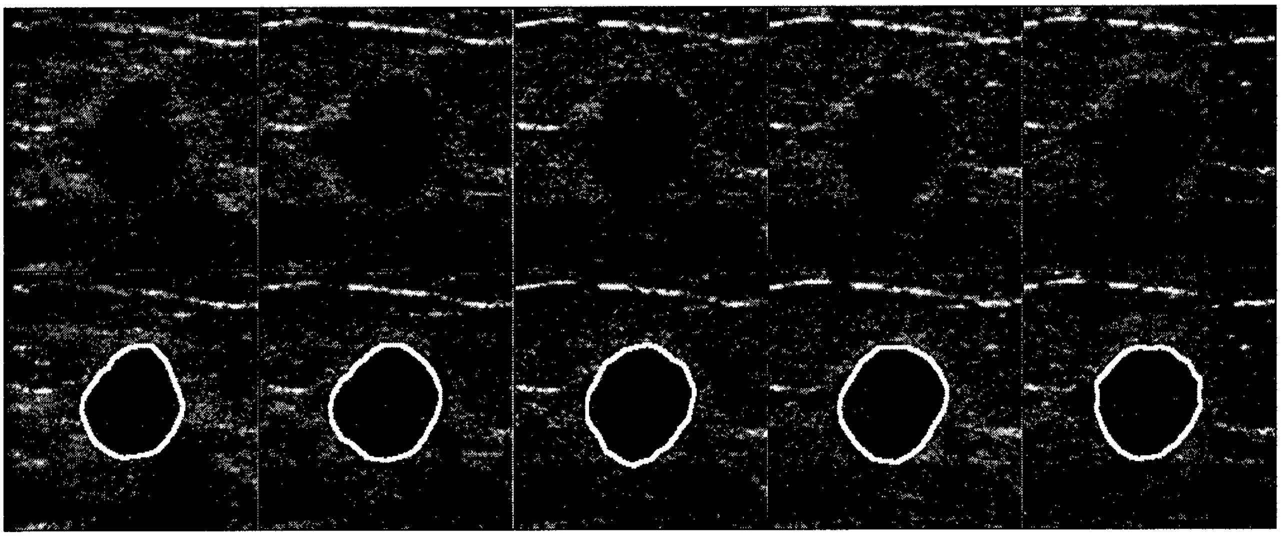


Figure 1: Five slices containing a malignant mass and the result of computer segmentation.

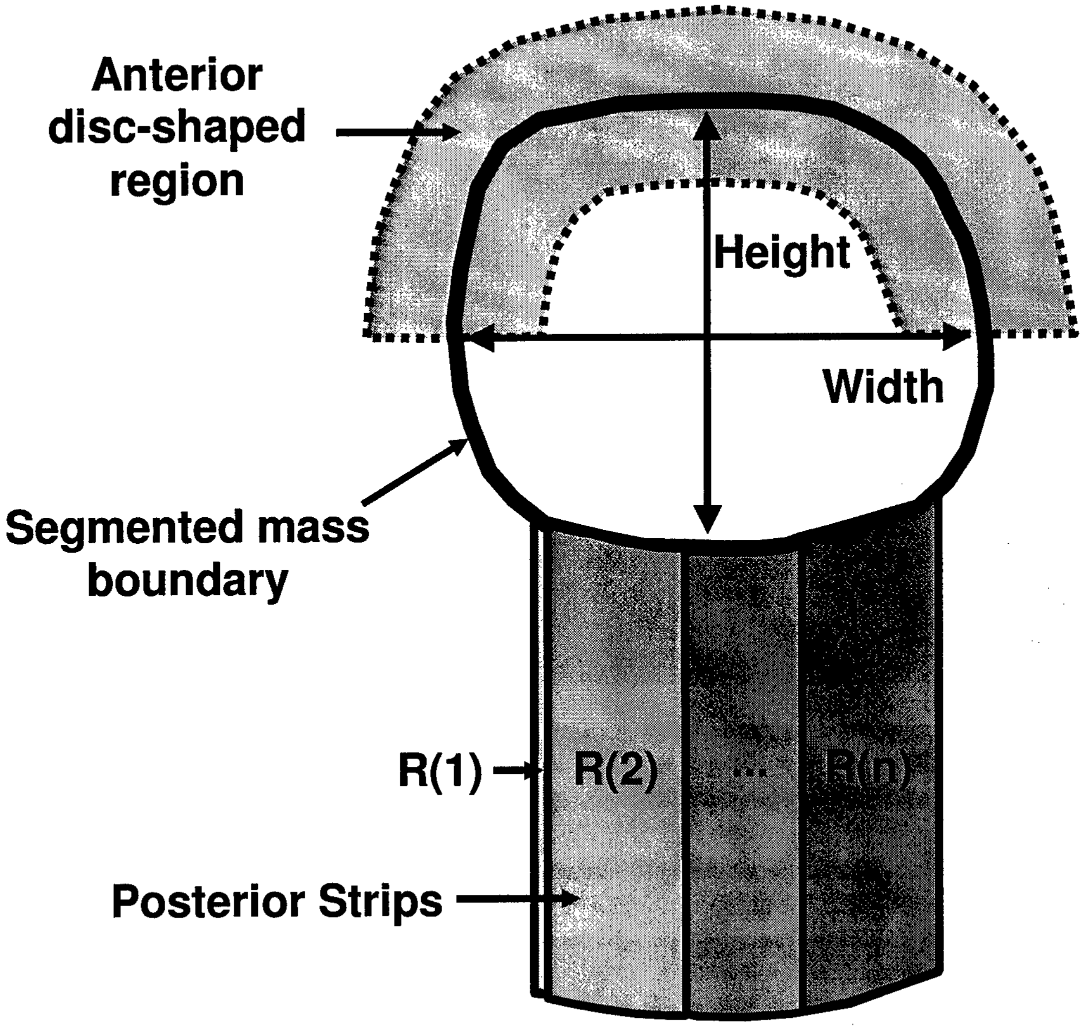
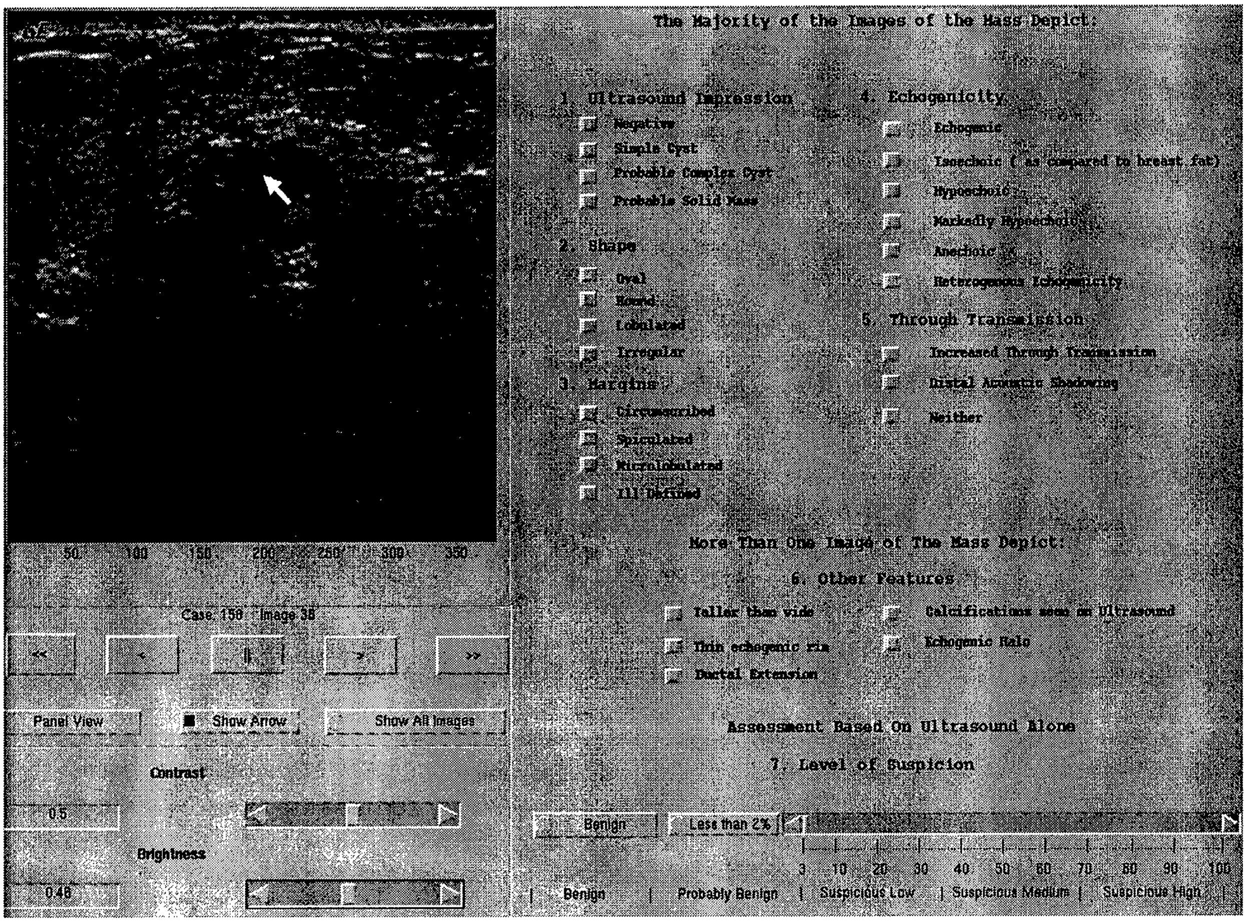


Figure 2: For feature extraction, the width *W* and height *H* of the mass on a slice were defined as the widest and the tallest cross-sections of the mass in that slice, respectively. The mean gray level values within the overlapping posterior strips *R(i)* and the segmented mass were used to define the posterior shadowing features. The disc-shaped regions for texture feature extraction followed the shape of the mass and contained partly the segmented mass and partly its margins. An example of the anterior disc-shaped region is shown as the gray area above the segmented mass.



The Majority of the Images of the Mass Depict:

1. Ultrasound Impression
□ Negative
□ Simple Cyst
□ Probable Complex Cyst
□ Probable Solid Mass

2. Shape
□ Oval
□ Round
□ Lobulated
□ Irregular

3. Margins
□ Circumscribed
□ Spiculated
□ Microlobulated
□ Ill Defined

4. Echogenicity
□ Echogenic
□ Isoechoic (as compared to breast fat)
□ Hypoechoic
□ Markedly Hypoechoic
□ Anechoic
□ Heterogeneous Echogenicity

5. Through Transmission
□ Increased Through Transmission
□ Distal Acoustic Shadowing
□ Neither

More Than One Image of The Mass Depict:

6. Other Features
□ Taller than wide
□ Thin echogenic rim
□ Ductal Extension
□ Calcifications seen on Ultrasound
□ Echogenic Halo

Assessment Based On Ultrasound Alone

7. Level of Suspicion

Benign | Less than 2%

Case: 150 Image 38

■ Show Arrow

Contrast 0.5

Brightness 0.48

Figure 3: The graphical user interface. The biopsy-proven lesion was marked by an arrow, which could be switched off when the radiologist assessed the mass. The interface allowed the users to navigate through the volume, and to adjust the contrast and brightness. The radiologists first provided their assessment for the mass in six categories, which were 1) overall US impression; 2) shape; 3) margins; 4) echogenicity; 5) through transmission; and 6) other features. They then provided a likelihood of malignancy rating without CAD. Finally, the computer's malignancy score for the mass was displayed and the radiologists had an option to revise their rating after taking into consideration the computer's opinion.

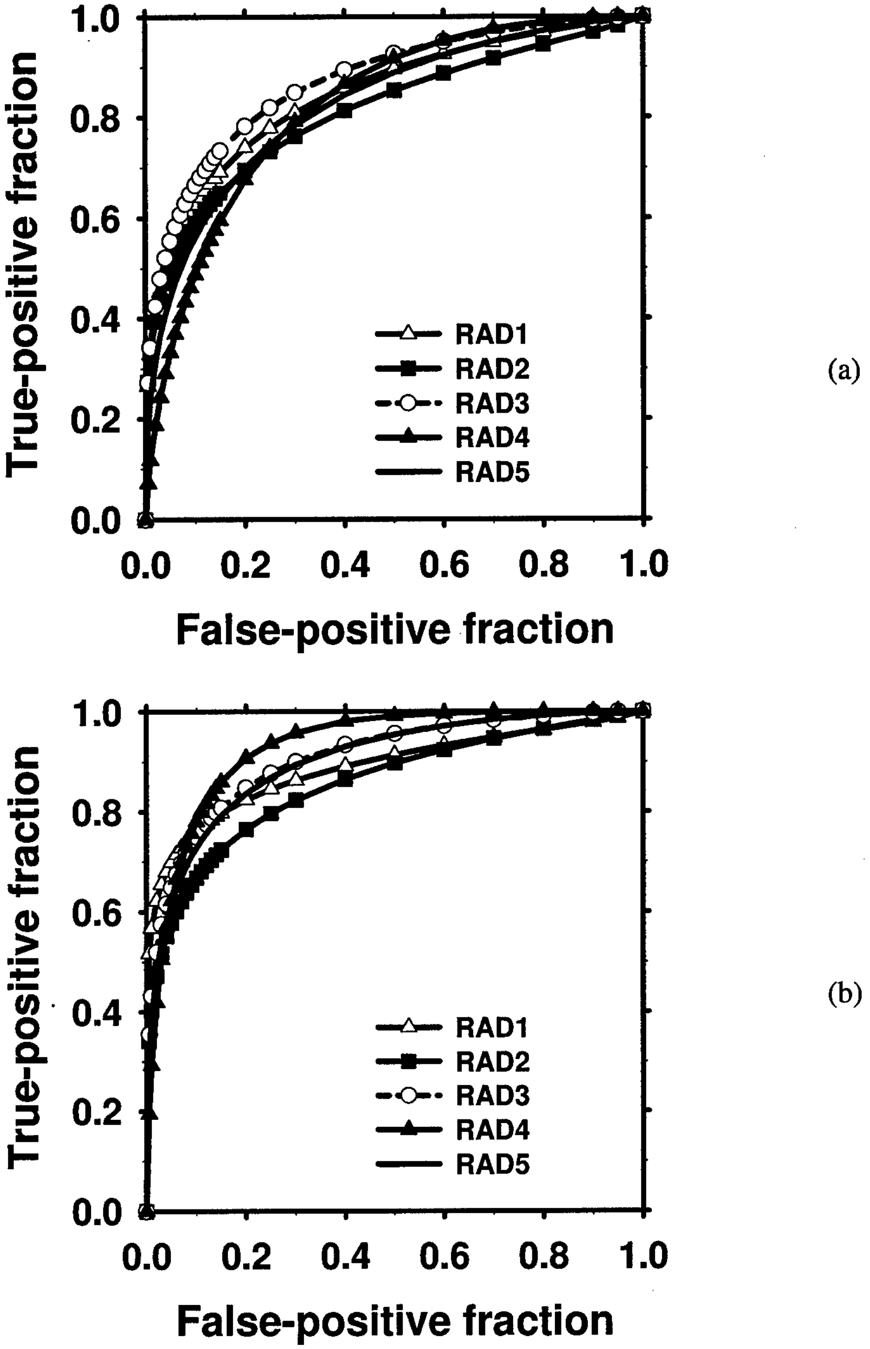


Figure 4: (a) The ROC curves of the five radiologists without CAD and (b) with CAD. The area under the ROC curve A_z and the partial area index above a sensitivity threshold of 0.9 $A_z^{(0.9)}$ are shown in Table 2 for each radiologist.

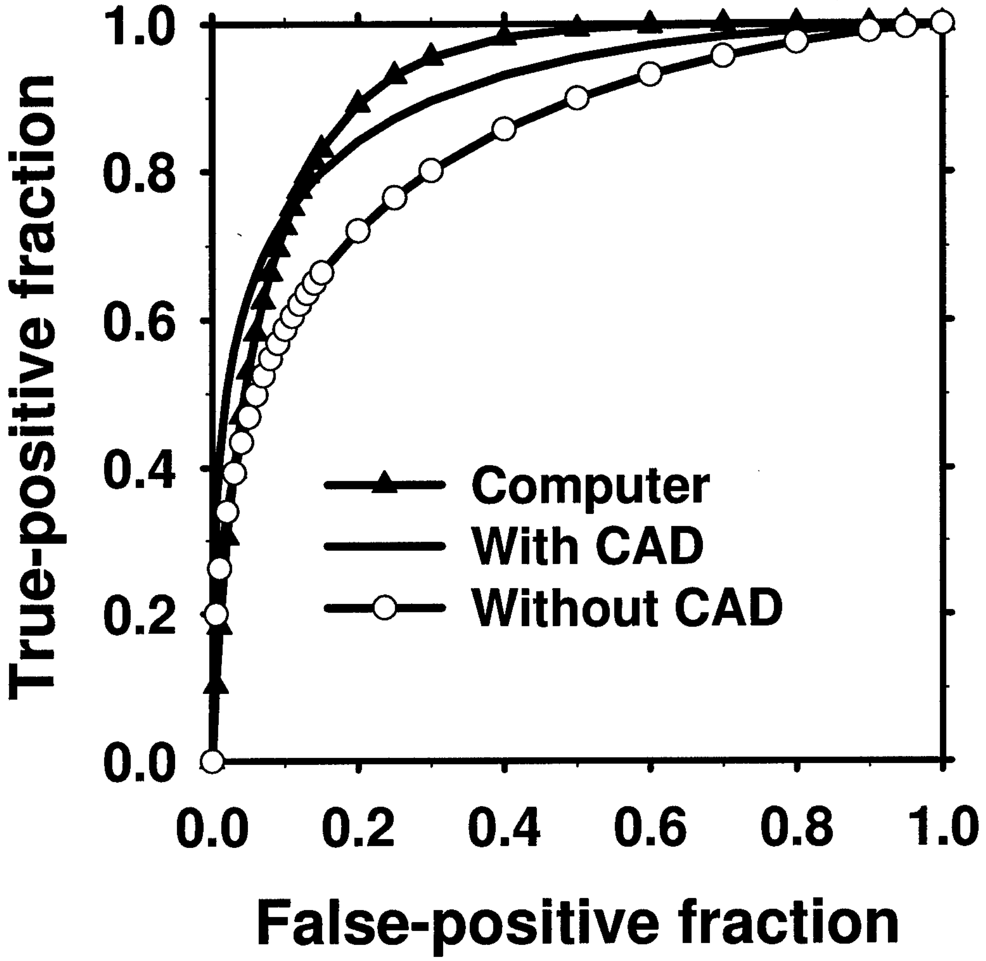


Figure 5: The average ROC curves of the radiologists with and without CAD, and the ROC curve of the computer classifier. The average ROC curves were constructed by using the mean a and b values of the individual observers' ROC curves shown in Figure 4.

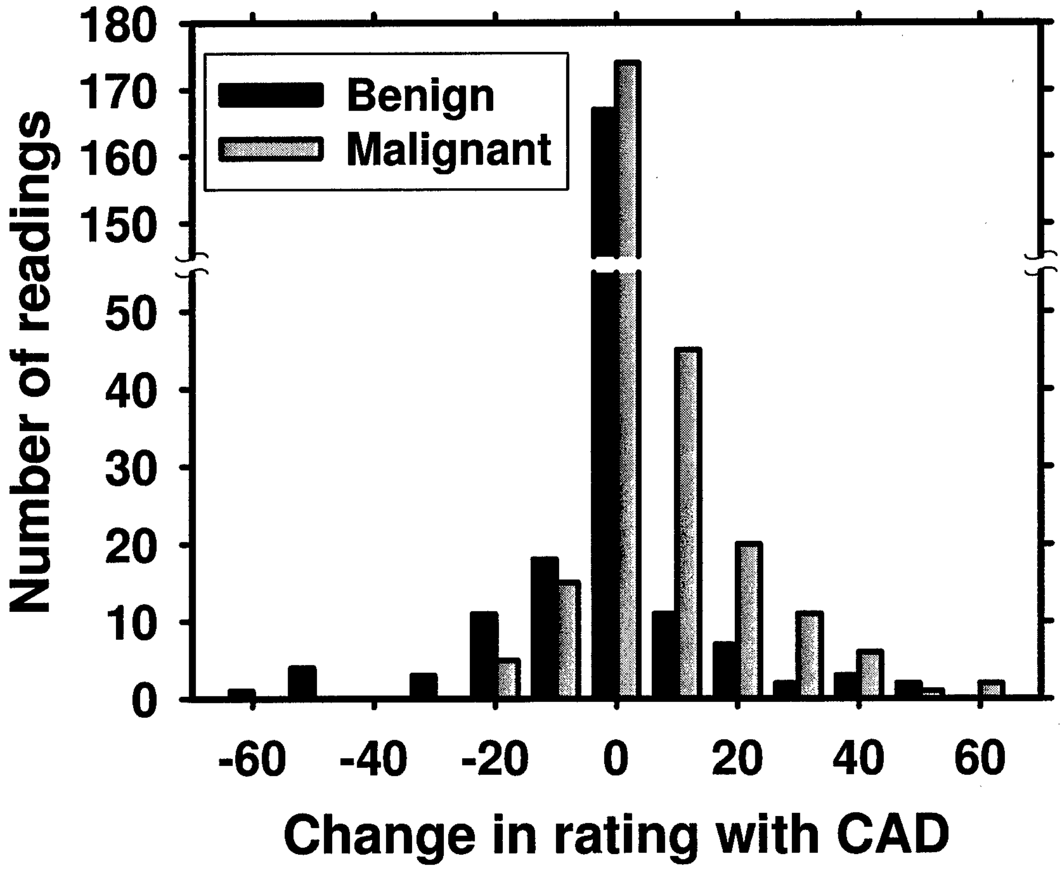


Figure 6: The histogram of the change in radiologists' ratings with CAD. For the majority of the masses (59% of malignant masses and 70% of benign masses) the change was in the range of -4 to 4. When the change in the scores with CAD was greater than or equal to the range of -5 to 5, the change was called substantial. For malignant masses, the ratings were substantially increased for an average of 34% (95/280) and decreased for 7% (19/280) of the readings. For benign masses, the ratings were substantially increased for 14% (32/230) and decreased for 17% (38/230) of the readings.

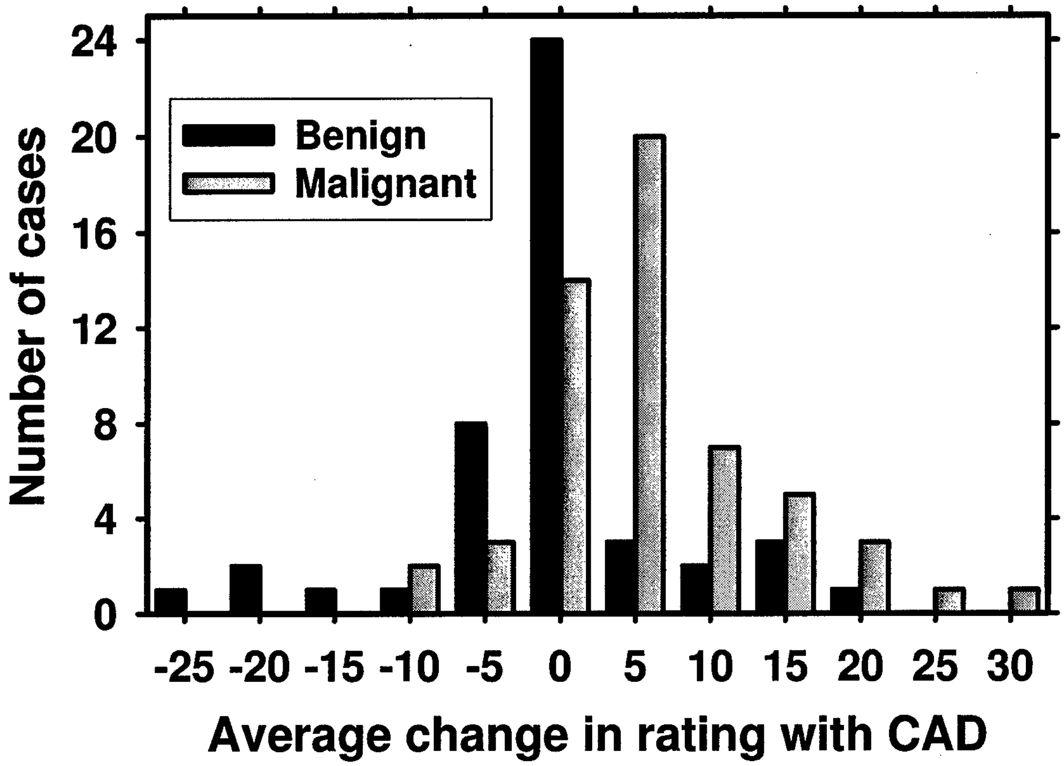


Figure 7: The histogram of the mean change in the LM ratings of radiologists with CAD. The mean change for a mass was computed by averaging the changes in the LM ratings for that mass over the five radiologists who participated in the study. For benign masses, the overall average LM rating decrease was 0.77, which did not achieve statistical significance (p=0.51). For malignant masses the overall average LM rating increase was 5.59, which was statistically significant (p<0.0001).

Appendix 4

Abstract Submission Preview

Abstract ID: 4412549
Submission Type: Scientific Papers

Berkman Sahiner
University of Michigan Medical Center
Phone: 734-647-7429
Fax: 734-615-5513
E-Mail: berki@umich.edu

THE EFFECT OF A MULTI-MODALITY COMPUTER CLASSIFIER ON RADIOLOGISTS' ACCURACY IN CHARACTERIZING BREAST MASSES USING MAMMOGRAMS AND VOLUMETRIC ULTRASOUND IMAGES: AN ROC STUDY

B Sahiner (P); H Chan; L M Hadjiiski; M A Roubidoux; C P Paramagul; M A Helvie ; et al.

PURPOSE

Computer-aided diagnosis (CAD) methods have previously been developed to assist radiologists in characterizing breast masses on mammograms and ultrasound (US) images. In this study, we developed a classifier that merged information from both modalities, and assessed its effect on radiologists' accuracy.

METHOD AND MATERIALS

The data set consisted of images from 67 patients containing biopsy-proven solid masses (32 benign and 35 malignant). An experienced radiologist identified the region of interest (ROI) containing the lesion on both modalities. The 3D US volumetric data were collected as cine-clips when the transducer was translated across the lesion. US and mammographic features were automatically extracted based on the margin, spiculation, shadowing, and shape characteristics of the masses. The features were combined into a malignancy score using a computer classifier designed with a leave-one-case-out method. Five MQSA radiologists participated in the ROC study. First, the radiologist read the mammogram ROIs, and provided a BIRADS score and a malignancy rating. Second, the US images were displayed along with the mammogram ROIs, the radiologist provided a second malignancy rating, and recommended: (i) 1-year follow-up; (ii) short-term follow-up; or (iii) biopsy. Third, the computer score was displayed, and the radiologist provided a third malignancy rating and revised the recommended action. The classification accuracy was quantified using the area under ROC curve, Az.

RESULTS

The computer classifier achieved a test Az value of 0.91. When reading mammograms alone, the radiologists had an average Az of 0.88 (range: 0.82-0.93). When the mammograms were supplemented by US images, the average Az increased to 0.92 (range:0.86-0.96). With CAD, the average Az increased significantly (p=0.03) to 0.95 (range:0.90-0.98). The average sensitivity for biopsy recommendation also improved from 0.96 to 0.98, and average specificity improved from 0.37 to 0.39.

CONCLUSIONS

The radiologists were more accurate in characterizing masses when both mammograms and volumetric US images were available. A well-trained computer algorithm can improve radiologists' accuracy even in this multi-modality reading condition.

Appendix 5

Fusion of mammographic and sonographic computer-extracted features for improved characterization of breast masses

Berkman Sahiner, Heang-Ping Chan, Lubomir M. Hadjiiski, Marilyn A. Roubidoux,

Chintana Paramagul, Mark A. Helvie, Gerald L. LeCarpentier

Department of Radiology, University of Michigan, Ann Arbor

Abstract

Ultrasound and mammography are two commonly used modalities for characterization of breast masses. Computerized classification methods have been developed for each modality for the purpose of aiding the radiologists in making a biopsy recommendation. Combining the diagnostic information from these two modalities may further increase the accuracy of computer classifiers for differentiation of malignant and benign masses. In this study, we developed a computerized multi-modality classifier that used computer-extracted information from 3D ultrasound images and digitized mammograms. Our data set contained mammograms and 3D ultrasound images from 67 patients. Thirty-two masses were benign and 35 were malignant. The feature space for our multi-modality classifier consisted of case-based ultrasound and case-based mammographic features. Ultrasound features were extracted based on the margin, shadowing, and shape characteristics of the masses that were segmented using an automated 3D algorithm. Features extracted from different ultrasound slices were averaged to yield case-based ultrasound features. Mammographic features were extracted based on texture, morphological, and spiculation characteristics. Features extracted from different views were averaged to yield case-based mammographic features. We used a leave-one-case-out resampling scheme for classifier design, which included both the feature selection and linear discriminant analysis stages. The area Az under the test receiver operating characteristic curve for the multi-modality classifier was 0.92. In comparison, the classifiers based on ultrasound and mammographic features alone had Az values of 0.88 and 0.86, respectively. For comparison with the multi-modality computer classifier, five experienced breast radiologists provided malignancy ratings based on the same mammograms and 3D ultrasound images. The radiologists' Az values ranged between 0.86 and 0.96 (average: 0.92). This study indicates that a multi-modality computer classifier can be designed for differentiation of malignant and benign breast masses which could achieve an accuracy comparable to that of experienced MQSA radiologists.

1. Introduction

A large percentage of breast biopsies are performed unnecessarily with an outcome of benign conditions (Bassett *et al.* 1992; Hermann *et al.* 1987). Computer-aided diagnosis (CAD) has a potential to assist the radiologists in reducing benign biopsies by providing them a consistent and reproducible second opinion. Computerized feature extraction and classification methods for characterization of breast masses on mammograms (Rangayyan *et al.* 1996; Sahiner *et al.* 1998; Huo *et al.* 1998) and on ultrasound (US) images (Chen *et al.* 1999; Horsch *et al.* 2002; Chen *et al.* 2003; Sahiner *et al.* 2004) have been active areas of research. Observer performance experiments indicate that the accuracy of radiologists' characterization of breast masses on mammograms (Chan *et al.*

1999; Huo *et al.* 2002) and on US images (Sahiner *et al.* 2003; Horsch *et al.* 2004) may be significantly improved if they are aided by a well-trained CAD system. The purpose of this study was to design a multi-modality classifier that uses features from both mammograms and US images with the purpose of further improving the radiologists' accuracy by CAD. The performance of the multi-modality classifier was compared to those of the single-modality computer classifiers and of experienced radiologists reviewing both modalities.

2. Methods

2.1 Feature extraction for classification of breast masses on mammograms

The first step in our feature extraction method was mass segmentation. Our automated mass segmentation method (Sahiner *et al.* 2001) was based on an active contour model that followed an initial segmentation using K-means clustering. Since the active contour model could not be used to segment spiculations, an additional segmentation stage was designed for spiculation detection (Sahiner *et al.* 2001).

After segmentation, morphological features were extracted from the segmented mass shapes. The extracted features included a Fourier descriptor, convexity, rectangularity, perimeter, contrast, circularity, perimeter-to-area ratio, area, normalized radial length (NRL) mean, NRL entropy, NRL area ratio, NRL standard deviation, and NRL zero crossing count. Three spiculation features were extracted from a spiculation measure defined for the pixels along the boundary of the mass (Hadjiiski *et al.* 2001). Run-length statistics (RLS) texture features were extracted from the band of pixels surrounding the mass after the band was transformed into Cartesian coordinates using the rubber-band straightening transform (Sahiner *et al.* 1998).

2.2 Feature extraction for classification of breast masses on 3D US volumes

The 3D US images were acquired using an experimental system previously developed and tested at our institution (Bhatti *et al.* 2001). A 3D active contour model initialized with a radiologist-defined 3D ellipsoid was used to segment the mass (Sahiner *et al.* 2004). After mass segmentation, morphological and texture features were extracted from each slice containing the mass. The morphological features included the width-to-height ratio and posterior shadowing features. The definitions of these features can be found in the literature (Sahiner *et al.* 2004). The texture features were extracted from SGLD matrices derived from 2D slices of the 3D data set. Since the margins of the mass contain the richest information for characterization, these features were extracted from two disk-shaped regions containing the mass boundary on the upper and lower margins of the mass. Six texture feature measures that are invariant under linear, invertible gray scale transformations were extracted. More details about texture features used in this study can be found in the literature (Sahiner *et al.* 2004).

2.3 Multi-modality classifier

The methods described above extracted features from each view for the mammograms, and each slice of the 3D volume for US images. We have previously studied different methods for combining these multi-modality features (Sahiner *et al.* 2004). In this study, we followed the following approach, which was found to be one of the successful strategies for multi-modality classifier design in our previous study: (i) averaged the feature vectors from each mammographic view to obtain a case-based mammographic feature vector (ii) averaged the feature vectors from each US slice to obtain a case-based US feature vector (iii) pooled case-based mammographic and US features in a combined feature space for classifier design, which included stepwise feature selection and linear discriminant analysis. A leave-one-case-out methodology was used to train and test the classifier with N=67 cases. The training included feature selection and the computation of classifier coefficients for the selected features using N-1 cases. The test scores were analyzed using receiver operating characteristic (ROC) methodology. The classification accuracies using single and multi-modality features were compared in terms of the area A_z under the ROC curve.

2.4 Data set

Our data set consisted of US volumes and mammograms from 67 patients who had a mammographically visible solid mass deemed suspicious or highly suggestive of malignancy. All patients underwent biopsy or fine needle aspiration. Thirty two of the masses were benign and 35 were malignant. The total number of mammographic views was 163, with each case containing between one and three views (CC, MLO, or LAT).

The biopsied mass on the mammograms and the US volumes was identified by an MQSA (Mammography Quality Standards Act) qualified radiologist using clinical images and case reports to confirm that the identified region contained the biopsied mass. Five radiologists read the mammograms and 3D US images on a high-quality computer monitor using a graphical user interface with which they could view the mammographic regions of interest, navigate through 3D volumes, adjust the window and level of the displayed images, and enter a malignancy rating between 1 and 100 (higher rating indicating higher likelihood of malignancy). All radiologists were MQSA qualified, and were either fellowship-trained in breast imaging or had over 25 years of experience in breast imaging. There was no time limit for the radiologists to read a case. The case reading order was randomized for each radiologist.

3. Results

The computer classifier using the US images alone, mammograms alone, and the combined feature space had A_z values of 0.88±0.04, 0.86±0.05, and 0.92±0.03, respectively. The ROC curves with the single-modality classifiers and the multi-modality classifier are shown in figure 1. Although the multi-modality classifier had higher accuracy, the difference between the A_z values of the multi-modality and single-modality classifiers did not reach statistical significance, probably because of the small sample size. The A_z values of the five radiologists ranged between 0.86 and 0.96. The A_z value of their average ROC curve, computed by averaging the a and b values in ROC analysis, was 0.92. Figure 2 compares the ROC curve of the computer classifier to that of the individual radiologists.

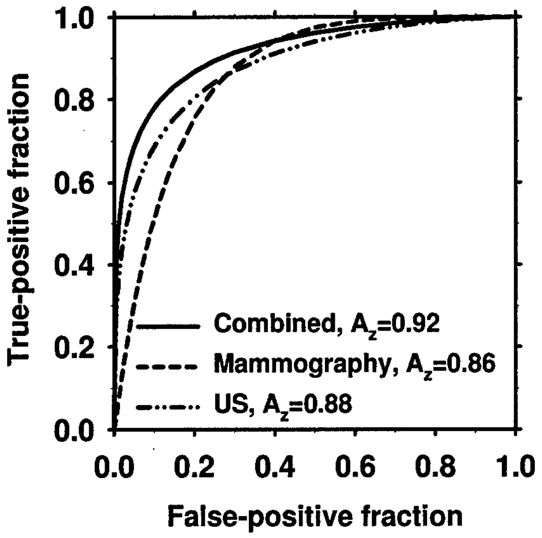


Figure 1: Comparison of the ROC curves for the single-modality and multi-modality classifiers.

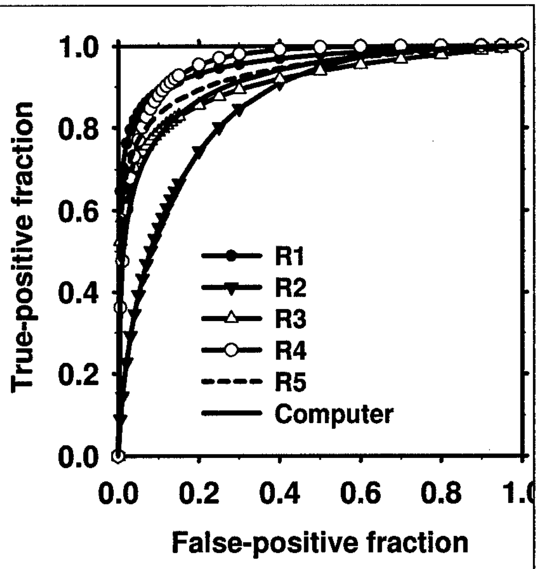


Figure 2: Comparison of the ROC curves of the five radiologists and the multi-modality computer classifier.

4. Conclusion

Our results indicate that a multi-modality classifier that combines computer-extracted features from mammograms and US images may improve the accuracy of the single-modality classifiers.

We plan to enlarge our data set to investigate if the observed improvement with multi-modality CAD is generalizable, and to test the statistical significance of the difference between the multi- and single-modality classifiers. We will also perform observer performance studies to investigate the effect of our multi-modality computer classifier on radiologists' accuracy in characterizing malignant and benign masses.

Acknowledgements

This work was supported in part by U.S. Army Medical Research Materiel Command grant DAMD17-01-1-0328 and by USPHS grants CA91713 and CA95153. The content

of this paper does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned should be inferred.

References

Bassett, L. W., T. H. Liu, A. I. Giuliano and R. H. Gold. 1992. The prevalence of carcinoma in palpable vs impalpable, mammographically detected lesions. *AJR* 158: 688-689.

Bhatti, P. T., G. L. LeCarpentier, M. A. Roubidoux, J. B. Fowlkes, M. A. Helvie and P. L. Carson. 2001. Discrimination of sonographically detected breast masses using frequency shift color Doppler imaging in combination with age and gray scale criteria. *Journal of Ultrasound in Medicine* 20: 343-350.

Chan, H.-P., B. Sahiner, M. A. Helvie, N. Petrick, M. A. Roubidoux, T. E. Wilson, D. D. Adler, C. Paramagul, J. S. Newman and S. S. Gopal. 1999. Improvement of radiologists' characterization of mammographic masses by computer-aided diagnosis: an ROC study. *Radiology* 212: 817-827.

Chen, C. M., Y. H. Chou, K. C. Han, G. S. Hung, C. M. Tiu, H. J. Chiou and S. Y. Chiou. 2003. Breast lesions on sonograms: computer-aided diagnosis with nearly setting-independent features and artificial neural networks. *Radiology* 226: 504-514.

Chen, D. R., R. F. Chang and Y. L. Huang. 1999. Computer-aided diagnosis applied to US of solid breast nodules by using neural networks. *Radiology* 213: 407-412.

Hadjiiski, L. M., B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie and M. N. Gurcan. 2001. Analysis of Temporal Change of Mammographic Features: Computer-Aided Classification of Malignant and Benign Breast Masses. *Med. Phys.* 28(11): 2309-2317.

Hermann, G., C. Janus, I. S. Schwartz, B. Krivisky, S. Bier and J. G. Rabinowitz. 1987. Nonpalpable breast lesions: Accuracy of prebiopsy mammographic diagnosis. *Radiology* 165: 323-326.

Horsch, K., M. L. Giger, L. A. Venta and C. J. Vyborny. 2002. Computerized diagnosis of breast lesions on ultrasound. *Med. Phys.* 29: 157-164.

Horsch, K., M. L. Giger, C. J. Vyborny and L. A. Venta. 2004. Performance of computer-aided diagnosis in the interpretation of lesions on breast sonography. *Acad. Radiol.* 11(3): 272-280.

Huo, Z. M., M. L. Giger, C. J. Vyborny and C. E. Metz. 2002. Breast cancer: Effectiveness of computer-aided diagnosis - Observer study with independent database of mammograms. *Radiology* 224(2): 560-568.

Huo, Z. M., M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt and K. Doi. 1998. Automated computerized classification of malignant and benign masses on digitized mammograms. *Acad. Radiol.* 5: 155-168.

Rangayyan, R. M., N. El-Faramawy, J. E. L. Desautels and O. A. Alim (1996). Discrimination between benign and malignant breast tumors using a region-based measure of edge profile acutance. *Digital Mammography '96* Eds. K. Doi, M. L. Giger, R. M. Nishikawa and R. A. Schmidt. Amsterdam, Elsevier. 213-218.

Sahiner, B., H. P. Chan, L. M. Hadjiiski, M. A. Roubidoux, C. Paramagul, M. A. Helvie and C. Zhou. 2004. Multi-modality CAD: Combination of computerized classification techniques based on mammograms and 3D ultrasound volumes for improved accuracy in breast mass characterization. *Proc. SPIE* 5370: 67-74.

Sahiner, B., H. P. Chan, N. Petrick, M. A. Helvie and M. M. Goodsitt. 1998. Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis. *Med. Phys.* 25: 516-526.

Sahiner, B., H.-P. Chan, N. Petrick, M. A. Helvie and L. M. Hadjiiski. 2001. Improvement of mammographic mass characterization using spiculation measures and morphological features. *Med. Phys.* 28: 1455-1465.

Sahiner, B., H. P. Chan, M. A. Roubidoux, M. A. Helvie, J. Bailey and L. M. Hadjiiski. 2003. An ROC study on characterization of malignant and benign breast masses in 3D ultrasound volumes: The effect of computer-aided diagnosis on radiologists' characterization accuracy. *RSNA 2003*, Chicago, Ill, Radiological Society of North America.

Sahiner, B., H. P. Chan, M. A. Roubidoux, M. A. Helvie, L. M. Hadjiiski, A. Ramachandran, G. L. LeCarpentier, A. Nees, C. Paramagul and C. E. Blane. 2004. Computerized characterization of breast masses on 3-D ultrasound volumes. *Med. Phys.* 31(4): 744-754.